FUJITSU

WHITE PAPER

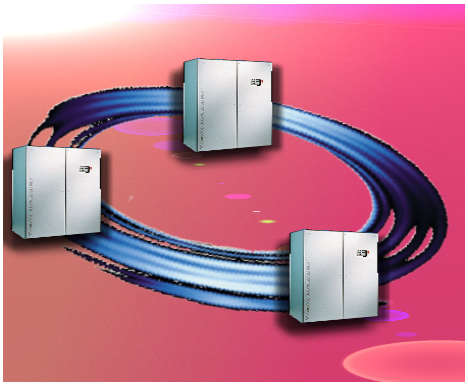
Katastrophenschutz-Konzepte in BS2000/OSD

Ausgabe April 2010

Seiten 45

Zusammenfassung

Katastrophenschutz-Konzepte werden benötigt, um im Falle einer Katastrophe betroffene Anwendungen und ihre Daten verfügbar zu machen und damit unterbrochene Geschäftsabläufe wieder aufsetzen zu können.



Die vorliegende Schrift bietet eine Einführung und Übersicht zum Thema Katastrophenschutz (Disaster Recovery) speziell für BS2000-Systeme unter Nutzung der Datenspiegelung auch über weite Entfernungen.

Sie spiegelt zu diesem wichtigen und hochaktuellen Thema die Erfahrung von Fujitsu und seiner Kunden wider und sie zeigt auch beim Katastrophenschutz die ausgereiften Eigenschaften von BS2000/OSD- und seiner Partnerprodukte.

Ziel ist es, dem verantwortlichen Rechenzentrums-Betreiber die von Fujitsu empfohlenen Varianten von Katastrophenschutz-Konfigurationen vorzustellen und die Voraussetzungen für eine KS-Konfiguration sowie die prinzipiellen Abläufe im Katastrophenfall (Failover und Failback) zu beschreiben. Dem Systemadministrator werden die technischen Möglichkeiten zur Katastrophen-vorsorge dargestellt.

Alle benötigten Techniken und Hilfsmittel

- **Symmetrix Remote Data Facility (SRDF™), Storage Host Component for BS2000/OSD (SHC-OSD)**
- **WDM-Technologie für synchrone Datenspiegelung über weite Entfernungen**
- **Einsatz von virtuellen Hosts (resp. DNS-Eingriffe) für Netzumschaltung im K-Fall**
- **HIPLEX AF für automatische Ausfallerkennung und Umschaltung im K-Fall**
- **Einsatz von Globalspeicher in KS-Konfigurationen**
- **Einsatz der Symmetrix-Funktion TimeFinder für asynchronen KS**

werden verständlich dargestellt.

Stand der Beschreibung: Januar 2006

Inhalt

1 Einleitung 3
1.1 Was verstehen wir unter Katastrophenschutz ? 3
1.2 Was verstehen wir (in Abgrenzung dazu) unter Hochverfügbarkeit? 4
1.3 Verflechtung von Hochverfügbarkeit (HV) und Katastrophenschutz (KS) 4
1.4 Ziel und Aufbau dieses Dokuments 5
2 Begriffe, Definitionen, Einführung 6
2.1 Katastrophenschutz-Konfiguration 6
2.2 Ausweich-RZ 6
2.3 Symmetrischer KS / Einseitiger KS 6
2.4 Cold-Standby / Hot-Standby / Backup 6
2.5 Synchrone und asynchrone KS-Konfigurationen 6
2.6 Failover / Failback 6
2.7 Manueller und automatischer Katastrophenschutz 7
2.8 Zur Symmetrix-Funktion SRDF 7
2.9 Zur Symmetrix-Funktion TimeFinder/Mirror 10
2.10 Globalspeicher (GS) 10
2.11 Konsistenz-Gruppen 10
2.12 HIPLEX MSCF 11
2.13 HIPLEX AF 11
2.14 Storage Host Component im BS2000 (SHC-OSD) 13
2.15 Dual Recording By Volume (DRV) 13
3 Voraussetzungen für Katastrophenschutz im BS2000 15
3.1 Organisatorische Maßnahmen 15
3.2 Aufbau eines Standby-RZ 16
3.3 Vorbereitung Datennetz 16
3.4 Verfahren zur Verfügbarkeit der Online-Daten im Standby-RZ 17
3.5 Entfernungsabhängige Einschränkungen 24
3.6 Administrative Vorbereitungen für die Katastrophenvorsorge 25
3.7 Datensicherungs-Konzept 26
3.8 Stufenweise Einrichtung der KS-Voraussetzungen 27
4 KS-Konfigurationen 28
4.1 Synchrone enge KS-Konfiguration (X-Link-Konf.) 28
4.2 Synchrone enge KS-Konfiguration mit GS 30
4.3 Synchrone lose KS-Konfiguration (U-Link-Konf.) 32
4.4 Asynchrone lose KS-Konfiguration 34
4.5 Kombinierte KS-Konfigurationen 36
4.6 Konfigurationen mit mehr als einem Symmetrix Subsystem 37
4.7 Trennung von Anwendungen durch VM2000 37
5 Abläufe beim Failover 39
5.1 Automatischer Failover mit HIPLEX AF 39
5.2 Manueller Failover 41
6 Abläufe beim Failback 43
6.1 Failback mit Hilfe von HIPLEX AF 43
6.2 Manueller Failback 43
7 Literatur und Online-Verweise 45

1 Einleitung

1.1 Was verstehen wir unter Katastrophenschutz ?

Im IT-Bereich ist Katastrophenschutz diejenige Vorsorge, welche nach einer teilweisen oder vollständigen Zerstörung einer Produktionsstätte - der IT-Infrastruktur „Rechenzentrum“ - die Wiederaufnahme der Produktion und damit der geschäftswichtigen Anwendungen ermöglicht.

Unter einer Katastrophe soll der Ausfall eines Rechenzentrums durch Stromausfall oder Zerstörung (Brand, Wassereinbruch, Explosion, Erdbeben, Sturm, Sabotage etc) oder etwas spezifischer der Ausfall eines Hosts und der räumlich in der Nähe aufgestellten Speicherperipherie oder auch nur von Teilen der Speicherperipherie, die aktuelle Produktionsdaten enthält, verstanden werden. Im Ergebnis handelt es sich bei einer so verstandenen Katastrophe um ein auf äußere Einwirkungen zurückgehendes Ereignis, das in seiner Wirkung zu einem Abbruch des Produktionsbetriebes führt, und welches zur Weiterführung des Produktionsbetriebes die Umschaltung aller erforderlichen Betriebsmittel auf die in einem Standby-RZ (Ausweich-RZ) vorgehaltenen analogen Betriebsmittel erzwingt. Standby-RZ heißt in der folgenden Beschreibung entweder ein räumlich entferntes RZ mit gleichwertiger Hardwareausstattung oder ein für die Anwendungen nötiges Hardware-Equipment (im wesentlichen Host, Plattenspeicher, Netzkomponenten), das zumindest durch eine Brandschutzmauer vom Produktiv-Host und dessen Plattenspeicher-Subsystemen getrennt ist und eine eigene Stromversorgung hat.

Der Eintritt einer Katastrophe wird im Folgenden auch kurz mit „K-Fall“ bezeichnet.

Die Anforderungen an ein Katastrophenschutz-Konzept liegen also in der Bereitstellung einer Konfiguration (HW + SW), die es ermöglicht, bei einem Gesamtausfall eines RZ den Produktionsbetrieb in einem Standby-RZ aufnehmen zu können (site failure recovery), und dabei folgende Randbedingungen so gut wie möglich einzuhalten:

- Sicherstellung eines konsistenten und aktuellen, zumindest möglichst aktuellen, Datenbestands der Produktionsanwendungen im Standby-RZ
- Unveränderte Zugriffsmöglichkeiten der Anwender auf den Datenbestand (im Standby-RZ) durch „Netzumschaltung“ im K-Fall
- Keine langen Ausfallzeiten bis zum Wiederanlauf der Anwendungen
- Zugriff auf die in der Vergangenheit archivierten Daten und Sicherstellung der Fortführung des bisherigen Datensicherungskonzeptes auch im Standby-RZ
- Unterstützung bei der Rückverlagerung der Anwendungen ins wiederhergestellte RZ (Failback)

Optionale Anforderungen an ein Katastrophenschutz-Konzept sind

- Automatische Ausfallerkennung und automatischer Wiederanlauf
- Unterstützung größerer Entfernungen zwischen Produktions- und Standby-RZ, die über einen Bereich von wenigen Kilometern hinausgehen

Geht man davon aus, dass ein Ausweich-RZ (Host, Online-Peripherie, Netzkomponenten etc.) vorhanden ist, muss ein solches Konzept die folgenden wesentlichen Fragen klären:

1) Wie kommen die Daten der Anwendungen ins Ausweich-RZ?

2) Welche netzseitigen Vorkehrungen sind erforderlich, damit die Anwender nach dem Wiederanlauf im Ausweich-RZ innerhalb eines angemessenen Zeitraumes wie gewohnt ihre Verfahren (die auf einem anderen Host laufen) nutzen können (im Idealfall ohne Eingriffe bei den Anwendern).

Die folgende Abbildung fasst diese Anforderungen als Grundvoraussetzungen des Katastrophenschutzes bildlich zusammen:

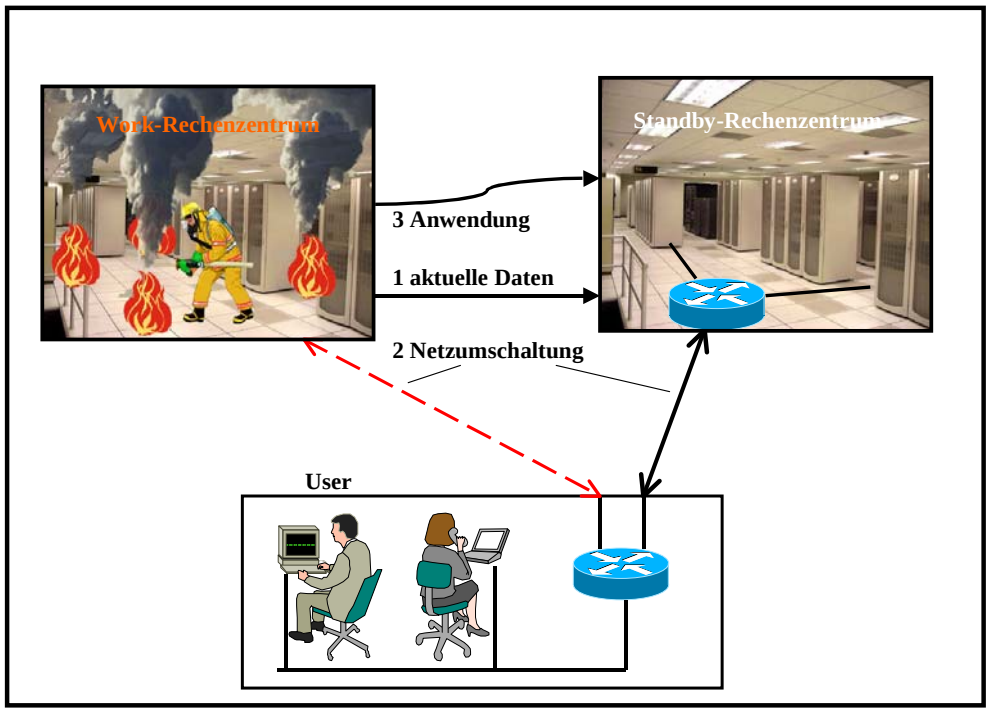


Abb. 1-1 Grundvoraussetzungen für KS

Die in der Literatur benutzten Begriffe

- Katastrophenschutz (KS)
- Katastrophenvorsorge
- Disaster Recovery (D/R)
- Disaster Protection
- Disaster Tolerance

werden als gleichbedeutend angesehen und in dieser Schrift synonym benutzt.

1.2 Was verstehen wir (in Abgrenzung dazu) unter Hochverfügbarkeit?

Eine Hochverfügbarkeits-Konfiguration dient in erster Linie dazu, Ausfälle einzelner Betriebsmittel (HW, aber auch SW) möglichst ohne Unterbrechung des Produktionsbetriebes zu überstehen (Minimierung von Ausfallzeiten, single failure recovery). Maßnahmen hierzu sind die hardwaremäßige Redundanz aller für die Aufrechterhaltung des Produktionsbetriebes notwendigen Betriebsmittel, die softwaremäßige Überwachung der Betriebsmittel, automatisierte Reaktionen auf Hard- und Softwarefehler. Der wesentliche Unterschied des Katastrophenschutzes gegenüber Hochverfügbarkeit besteht darin, dass die für die Aufrechterhaltung des Produktionsbetriebes redundanten Betriebsmittel und Daten räumlich entfernt sind und damit gegenüber zerstörenden Einwirkungen am Produktionsort geschützt sind.

1.3 Verflechtung von Hochverfügbarkeit (HV) und Katastrophenschutz (KS)

Im Idealfall sind dies zwei Ecksteine einer IT-Landschaft mit HV als Basis und KS als Ergänzung. Es ist nicht richtig, dass KS notwendig auch ein Maximum an HV voraussetzt oder impliziert: So gibt es KS-Verfahren, die zugunsten der hundertprozentigen Datenintegrität, Datenkonsistenz und Datenaktualität (im Standby-RZ) auch bei redundant vorgehaltener HW (wie etwa zwei über SRDF verbundener Symmetrix Systeme) bei Ausfall des einen Systems (oder der Verbindungen zwischen den Systemen) nicht mit dem verbleibenden System weiterarbeiten, sondern sicherheitshalber mit Fehlerbedingungen den Produktionsbetrieb beenden und erst nach Beurteilung des Ausfallszenarios durch hierfür ausgebildetes Personal ggf. den Betrieb auch mit nur einem System wiederaufnehmen (Stichwort „Domino-Modus“, s.2.8.2).

Somit sind HV und KS zunächst zwei verschiedene Ziele, deren individuelle Ausgestaltung aufeinander abgestimmt werden kann und muss. Die Risiken von Ausfallzeiten versus Datenverlust müssen kundenspezifisch bewertet werden, um eine optimale Abstimmung zu erreichen. Im Kapitel 4 dieses Dokuments werden Katastrophenschutz-Konfigurationen für BS2000-Systeme beschrieben. Die „synchrone enge Konfiguration“ beispielsweise bietet die Möglichkeit, Katastrophenschutz und Hochverfügbarkeit optimal zu verbinden und durch eine gemeinsame Software (HIPLEX AF) zu steuern. Wenn aufgrund von großen Entfernungen, Performance-Anforderungen oder aus Kostengründen diese Konfiguration nicht in Frage kommt, wird es leichte Abstriche entweder beim Katastrophenschutz oder bei der Hochverfügbarkeit geben.

1.4 Ziel und Aufbau dieses Dokuments

Die vorliegende Schrift bietet eine Einführung und Übersicht zum Thema Katastrophenschutz speziell für BS2000-Systeme unter Nutzung der SRDF-Funktion angeschlossener Symmetrix Systeme.

Ziel ist es, dem verantwortlichen Rechenzentrums-Betreiber die von Fujitsu Technology Solutions empfohlenen Varianten von Katastrophenschutz-Konfigurationen vorzustellen und die Voraussetzungen für eine KS-Konfiguration sowie die prinzipiellen Abläufe im Katastrophenfall (Failover und Failback) zu beschreiben. Dem Systemadministrator werden die technischen Möglichkeiten zur Katastrophenvorsorge dargestellt.

Alle benötigten Techniken und Hilfsmittel

- Symmetrix Remote Data Facility (SRDF), Symmetrix Host Component for BS2000/OSD (SHC-OSD)
- WDM-Technologie für synchrone Datenspiegelung über weite Entfernungen
- Einsatz von virtuellen Hosts (resp. DNS-Eingriffe) für eine Netzumschaltung im K-Fall
- HIPLEX AF für automatische Ausfallerkennung und automatische Umschaltung im K-Fall
- Einsatz von Globalspeicher in KS-Konfigurationen
- Einsatz der Symmetrix-Funktion TimeFinder/Mirror für asynchronen KS

werden verständlich dargestellt.

Für die detaillierte Planung und Realisierung einer kompletten Katastrophenvorsorge-Lösung sind viele Faktoren zu berücksichtigen. Relevante Komplexe sind z.B. die Anforderungen an die Datenaktualität im K-Fall, die Wahl einer Cold-Standby oder Hot-Standby Lösung, die vorhandene RZ-Architektur (Hosts, Online-Peripherie, Netz, VMs, Nearline-Peripherie), natürlich die zu schützenden Anwendungen und deren IO-Last, die RZ-Standorte, ihre Entfernungen, der Standort des Bandarchivs und die eingesetzten Daten-Sicherungskonzepte sowie die kundenspezifischen organisatorischen und personellen Randbedingungen, und schließlich die Entscheidung für eine manuelle oder automatische Ausfallerkennung sowie für eine manuelle, halbautomatische oder automatische Umschaltung im K-Fall.

Die Vielfalt der Einflussfaktoren begründet die Tatsache, dass keine schlüsselfertigen KS-Produktlösungen beschrieben werden können – vielmehr bietet das vorliegende Dokument eine Hilfestellung für eine gemeinsam mit einem unserer Competence Center zu erarbeitende Individuallösung zur Katastrophenvorsorge, deren Realisierung dann ebenfalls von Fujitsu Technology Solutions Fachkräften durchgeführt werden kann.

Bereits seit längerem haben viele unserer Kunden mit den vorhandenen Mitteln ihre Rechenzentren für Katastrophenschutz vorbereitet und Konfigurationen in eigener Regie oder mit Hilfe unserer Competence Center realisiert, die geeignet sind, in einem K-Fall die wichtigen geschäftlichen Daten zu retten und die kritischen IT-gestützten Verfahren schnell wieder zum Ablauf bringen zu können. Diese Schrift gibt eine Gesamt-Übersicht zum Thema Katastrophenschutz im BS2000/OSD und nennt die aktuellen Hilfsmittel und Verfahren, die von Fujitsu Technology Solutions zur Katastrophenvorsorge zur Verfügung gestellt werden.

Im folgenden Kapitel 2 des Dokuments werden Begriffe und Produkte erläutert, die für KS-Konzepte im BS2000 wichtig sind; die Kapitel über Produkte sind nur für den nicht damit vertrauten Leser gedacht.

Kapitel 3 beschreibt die oben bereits genannten Voraussetzungen für KS-Konzepte (Datenspiegelung, Netzanbindung) sowie ggf. zusätzlich erforderliche organisatorische und administrative Maßnahmen. Weiter werden Verfahren zur Verfügbarkeit der Online-Daten im Ausweich-RZ und Techniken zur Überbrückung weiter Entfernungen für Datenspiegelungen diskutiert.

Im Kapitel 4 werden die Konfigurationen beschrieben, die aus Sicht der Autoren für Katastrophenschutz-Konzepte in BS2000/OSD zu empfehlen sind. Diese unterscheiden sich durch Faktoren wie Distanz der RZs, Datenaktualität, Performance und die Möglichkeiten bei der Automatisierung.

Die Kapitel 5 und 6 beschreiben skizzenhaft die nötigen Vorgänge zur Wiederaufnahme des Betriebs im Standby-RZ bei einem K-Fall sowie die Rückumschaltung auf das Work-RZ.

2 Begriffe, Definitionen, Einführung

Ein zentrales Thema in diesem und auch den folgenden Kapiteln ist die Spiegelung der Daten zum Standby- bzw. Ausweich-RZ. Für Katastrophenschutz-Konzepte in BS2000/OSD empfehlen wir die hochwertigen Symmetrix-Plattensubsysteme von EMC^2 und die zugehörige Remote-Copy-Funktion SRDF und beschränken uns in diesem Dokument auf die Beschreibung dieser Systeme. Die Alternative der Hostgesteuerten Spiegelung über das BS2000-Subsysten DRV wird im Abschnitt 2.15 zusätzlich skizziert. Zum heutigen Zeitpunkt ist die Nutzung der FibreCAT CX Plattensubsysteme von EMC^2 und der zugehörigen Remote-Copy-Funktion Mirrorview auf nicht automatisierte Cold-Standby Katastrophenschutz-Konzepte in BS2000/OSD beschränkt und wird im Folgenden hier nicht weiter thematisiert.

2.1 Katastrophenschutz-Konfiguration

Eine Katastrophenschutz-Konfiguration (kurz KS-Konfiguration, im angelsächsischen Disaster Tolerant Architecture) ist eine Cluster-Architektur zweier räumlich getrennter Rechenzentren, für die spezielle administrative Maßnahmen zur Katastrophenvorsorge definiert sind und durchgeführt werden. Die Summe der Maßnahmen an HW, SW und Administration muss geeignet sein, ausgewählte relevante Anwendungen im Falle von Katastrophen-Situationen, die ein produktives RZ (in dem geschäftskritische Anwendungen laufen) tangieren, innerhalb eines akzeptablen Zeitraums in einem zweiten RZ zum Ablauf zu bringen. Dieser Zeitraum ist vom Kunden zu beziffern und die Konfiguration daraufhin abzustimmen.

2.2 Ausweich-RZ

Bilden zwei räumlich getrennte RZs zusammen eine Katastrophenschutz-Konfiguration, und in einem der beiden RZs tritt ein K-Fall ein, so wird das jeweils andere, intakt gebliebene RZ das **Ausweich-RZ** genannt.

2.3 Symmetrischer KS / Einseitiger KS

In einer KS-Konfiguration mit zwei getrennten Rechenzentren RZ1 und RZ2 sollen im K-Fall ein oder mehrere Server im Ausweich-RZ RZ2 die Anwendungen der Server im Produktions-RZ RZ1 übernehmen. Wenn im RZ2 ebenfalls produktive Anwendungen laufen, kann ebenso für diese eine Übernahme in das RZ1 eingeplant werden. Bei einem K-Fall im RZ2 wäre dann RZ1 das Ausweich-RZ. Die Rollen von Produktions-RZ und Standby-RZ sind dann vertauschbar, und es liegt eine symmetrische KS-Konfiguration vor. Laufen im RZ2 keine relevanten Anwendungen, wird man dafür auch kein K-Fall-Konzept benötigen. Dann ist das RZ2 immer das Ausweich RZ, und es liegt eine einseitige KS-Konfiguration mit Produktions-RZ RZ1 und Standby-RZ RZ2 vor.
Praxis ist auch, dass sich zwei Rechenzentren unterschiedlicher Unternehmen zusammenschließen und Failover-Ressourcen für die jeweils „fremden“ Anwendungen bereitstellen, wenn dies organisatorisch machbar ist.

2.4 Cold-Standby / Hot-Standby / Backup

Bezeichnungen für die Rolle eines Standby-RZ. Cold-Standby (oder Backup-RZ) bezeichnet ein Standby-RZ, das nicht produktiv genutzt und erst im K-Fall aktiviert wird (bestenfalls wird es für Testanwendungen genutzt). Bei einem Cold-Standby kann es sich also nicht um eine symmetrische KS-Konfiguration handeln.
Hot-Standby ist ein laufendes und entsprechend vorbereitetes System, das ständig bereit ist, die Anwendungen des Produktions-RZ im K-Fall zu übernehmen. Die Ausfallzeiten sind deshalb kürzer und es ist sowohl symmetrischer als auch einseitiger KS möglich.

2.5 Synchrone und asynchrone KS-Konfigurationen

Bei einer **synchronen KS-Konfiguration** wird exakte Aktualität des kritischen Datenbestands im Standby-RZ nach einer Katastrophe gefordert. Liegen im Ausweich-RZ stets die gleichen Online-Daten wie im Original-RZ vor, so kann in einem K-Fall im Ausweich-RZ ohne Datenverlust weitergearbeitet werden. Eine Konfiguration der beiden RZ, insbesondere der Plattenspeichersysteme und ihrer Remote-Copy-Funktion, die das ermöglicht, nennen wir eine synchrone KS-Konfiguration.

Liegen im Ausweich-RZ nur Daten des Original-RZ eines früheren Zeitpunkts (z.B. des Vortags) vor, so kann in einem K-Fall im Ausweich-RZ nur mit dem Verzicht auf die nach dem Erstellen der Datensicherung noch neu erstellten Datensätze (z.B. aller neu eingebrachten Datensätze desjenigen Tages, an dem der K-Fall eingetreten ist) weitergearbeitet werden. Eine solche Konfiguration der beiden RZ, insbesondere der Plattenspeichersysteme, nennen wir eine **asynchrone KS-Konfiguration**.
Bei einer asynchronen Katastrophenschutz-Konfiguration ist die Datenaktualität des Datenbestands im Ausweich-RZ nach einer Katastrophe geringer: Im Ausweich-RZ kann jederzeit ein Datenbestand reaktiviert werden, der dem letzten Konsistenzpunkt der Anwendungen entspricht – in regelmäßigen Abständen wird ein solcher Konsistenzpunkt im Standby-RZ durch Aktionen am Produktions-Host erzeugt (z.B. im Backup-Modus einer Oracle-Datenbank) und bis zum Erstellen des nächsten Konsistenzpunkts eingefroren. Auch für eine asynchrone KS-Konfiguration ist es erforderlich, vom Produktions-Host erzeugte Daten in ein Ausweich-RZ zu transferieren – hierfür kommen asynchrone Datenspiegelungen, periodische Übermittlungen von Daten mittels Filetransfer oder physikalischer Transport von Datenträgern in Frage.

2.6 Failover / Failback

In einem K-Fall sollen im Ausweich-RZ die Produktions-Anwendungen des ausgefallenen RZ zum Wiederanlauf gebracht werden. Dieser Vorgang und die Summe aller dazu notwendigen Aktionen werden mit **Failover** bezeichnet. Der Failover kann entweder durch manuelle oder auch durch automatische Aktionen durchgeführt werden.

Nach einem Failover und der Wiederherstellung des ausgefallenen Produktions-RZ soll es in einem Katastrophenschutz-Konzept weiterhin möglich sein, eine Rückverlagerung der Anwendungen auf die ursprünglichen Betriebsmittel vorzunehmen. Dieser Vorgang und die Summe aller dazu notwendigen Aktionen werden mit **Failback** bezeichnet.

2.7 Manueller und automatischer Katastrophenschutz

Unter automatischem Katastrophenschutz ist zu verstehen, dass der Ausfall von mehreren Komponenten oder sogar eines gesamten RZ von einer Überwachungsssoftware, bei BS2000-Systemen ist dies HIPLEX AF, erkannt und automatisch behandelt wird. Es werden, falls noch möglich, alle SW-Ressourcen einer Anwendung beendet und dann am Standby-Host mit Spiegelplatten neu gestartet.

Wenn lediglich die HW- und System-Voraussetzungen geschaffen werden, um eine Anwendung am Ausweich-RZ ablaufbar und bedienbar zu machen, die Anwendung aber nach einem Ausfall durch einen Mitarbeiter neu gestartet werden muss (inkl. Herstellung der erforderlichen Ablauf-Umgebung), bezeichnen wir dies als manuellen Katastrophenschutz.

Eine Zwischenlösung ist ein halbautomatischer Katastrophenschutz, bei dem die Überwachungssoftware zwar den K-Fall anhand vordefinierter Kriterien erkennt und meldet – also die Entscheidung erleichtert, aber keine Umschaltung (Failover) einleitet. Dies erfordert dann, dass Personal vor Ort ist, oder automatisch benachrichtigt wird, um die Entscheidung über die Umschaltung anhand einer Reihe von Überprüfungen einzuleiten, hat aber den Vorteil, dass man einen unnötigen Failover verhindert, weil es sich evtl. nur um einen kurzzeitigen oder leicht zu behebenden Ausfall handelt. Für den Failover steht dann auch der gleiche automatisierte Ablauf wie beim automatischen Katastrophenschutz in Form von HIPLEX AF Switch-Units zur Verfügung, also ein „Failover auf Knopfdruck“.

Vorteile des automatisierten KS:
- Kein Personalaufwand im K-Fall (idealerweise)
- Vermeiden von menschlichen Fehlern in einer kritischen Situation
- Die Koordinierung des K-Fall-Ablaufs ist in der Überwachungssoftware implementiert.

Allerdings sind im K-Fall auch Mehrfachfehler (z.B. Rolling Disaster) denkbar, bei denen auch im automatisierten Ablauf eine Entscheidung eines Systemverwalters, also menschliches Eingreifen nötig ist.

Wichtig beim manuellen, wie aber auch beim automatischen KS ist es, dass die Abläufe im K-Fall detailliert dokumentiert sind und in regelmäßigen Abständen (z.B. ½ jährlich) getestet werden. Außerdem sollte das Know How für die einzelnen Abläufe nicht auf einzelne Mitarbeiter konzentriert sein.

2.8 Zur Symmetrix-Funktion SRDF

2.8.1 SRDF, SRDF-Modi

Die EMC-Funktionalität SRDF (Symmetrix Remote Data Facility) unterstützt die Datenspiegelung eines lokalen Symmetrix-Systems auf ein entferntes Symmetrix-System. Die beiden Symmetrix-Systeme sind dabei über mindestens zwei Remote Link Directors miteinander verbunden. Jede Symmetrix kann über Remote-Verbindungen mit maximal vier anderen verbunden sein. Unabhängig von der Entfernung kann SRDF in uni- oder bidirektionalen Konfigurationen verwendet werden. Bei Ausfall einer Symmetrix sind die aktuellen Daten auch immer in der entfernten Symmetrix vorhanden. Dadurch müssen die Daten nach einem Ausfall nicht erst wieder eingespielt, oder ggf. festgestellt werden, dass die Sicherungen unbrauchbar oder inkonsistent sind, und auch nicht auf einen veralteten Stand zurückgegangen werden. Ein über SRDF gespiegeltes Volume besteht aus der Source Unit (Original) und der Target Unit (Kopie der Daten), die über einen Remote Link verbunden sind. Die Source Unit liegt in der Symmetrix, die die Schreibaufträge im normalen SRDF-Betrieb erhält. Sie sendet die aktualisierten Daten zur Remote-Symmetrix, die die Daten zunächst im Cache und dann (asynchron) auf der Target Unit speichert. Ein SRDF-Paar wird vom Service-Techniker eingerichtet.

Es gibt die drei unterschiedlichen SRDF-Modi
- synchroner Modus
- semi-synchroner Modus (ab Microcode-Version e5771 nicht mehr unterstützt)
- Adaptive Copy Modus (asynchroner Modus)

Synchroner Modus bedeutet, dass Schreib-IOs auf eine SRDF-Platte erst dann dem Betriebssystem als erfolgreich beendet gemeldet werden, wenn die Daten in den lokalen Cache und auch über die SRDF-Verbindung in den Cache der Remote-Symmetrix geschrieben worden sind. Erst danach wird der nächste Schreibauftrag auf dieselbe Platte angenommen. Dadurch ist gewährleistet, dass die Daten aller ausgeführten Schreibaufträge nach dem Ausfall einer Symmetrix bzw. eines RZ noch verfügbar sind.

Im Unterschied dazu wird beim **semi-synchronen Modus** eine Schreib-IO als beendet gemeldet, sobald die Daten in den Cache der lokalen Symmetrix geschrieben sind. Der Abgleich des Cache in der Remote-Symmetrix erfolgt nur in bestimmten Intervallen, jedoch spätestens, sobald ein erneuter Schreibauftrag für dieselbe Platte eintrifft. D.h. bei Ausfall der lokalen Symmetrix kann pro Platte **ein** Schreibauftrag verloren sein. Dafür ist dieser Modus, gerade bei größeren SRDF-Entfernungen, performanter.

Beim **Adaptive Copy Modus** erfolgt der Abgleich der Daten zwischen den Steuerungen niederprior, sowohl verzögert zur Schreib-IO als auch in einer Reihenfolge, die nicht den Originär-IOs entspricht. Das bedeutet, dass bei Ausfall einer Steuerung der gespiegelte Datenbestand möglicherweise inkonsistent ist. Dieser Modus sollte daher nur dann zum Katastrophenschutz verwendet werden, wenn auch regelmäßige konsistente Datenstände erstellt und beispielsweise auf BCVs gesichert werden,

wie im Kapitel 4.4 beschrieben. Er ist nur dann einzusetzen, wenn der Verlust von Daten zwischen diesen Sicherungszeitpunkten im K-Fall tolerierbar ist.

Folgende Implikationen ergeben sich durch die Wahl dieser Modi:

Der synchrone Modus garantiert stets Gleichstand aller gespiegelten Daten, solange logisch zusammen-gehörige Daten nicht über SRDF Units auf verschiedenen Symmetrix-Steuerungen verstreut sind. Damit ist dieser Modus für den Katastrophenschutz bestens geeignet. Zu entscheiden ist dabei, ob die entfernungsabhängige (und ggf. Netztopologie-abhängige) Performance-Einbuße tolerierbar ist.

Der semi-synchrone Modus garantiert für die Daten einer Symmetrix-Steuerung Gleichstand bis auf maximal eine Schreib-IO pro Volume auf der entfernten Symmetrix-Steuerung sowie Datenkonsistenz in dem Sinne, dass Schreib-IOs auf verschiedene Volumes derselben Steuerung in der gleichen Reihenfolge auf die Target Units geschrieben werden, in der sie auch von der Anwendung geschrieben werden. Ist der Verlust von maximal einer Schreib-IO pro Volume im K-Fall aus Anwendersicht tolerierbar, so ist auch dieser Modus für den Katastrophenschutz geeignet, sofern beim Spiegeln mehrerer Symmetrix Systeme darauf geachtet wird, dass logisch zusammengehörige Daten nicht auf verschiedenen Symmetrix-Steuerungen verstreut sind. In der Regel verlieren Anwendungen, deren Daten in diesem Modus gespiegelt werden, keine Performance gegenüber dem ungespiegelten Modus, jedoch ist bei schreibintensiven Anwendungen die Performance-Ein-buße vergleichbar mit dem synchronen Modus.

Der Adaptive Copy Modus eignet sich, wegen der fehlenden Garantien, für die Datenkonsistenz im obigen Sinne einerseits und für den Gleichstand aller gespiegelten Daten zu jedem Zeitpunkt andererseits, nur bedingt zum Katastrophenschutz kritischer Anwendungen: der Datenbestand auf den Target Units im Standby-RZ nach einem K-Fall ist u.U. inkonsistent – aber durch ein regelmäßiges Erzeugen von Konsistenzständen, die im Standby-RZ auf additiven Spiegeln (z.B. BCVs) eingefroren werden, ist er bei geringeren Anforderungen an die Daten-Aktualität wegen seiner nur minimalen Auswirkungen auf die Gesamtperformance eine diskutable Alternative.

Schlagwortartig zusammengefasst:

SRDF-Modus	Datenverlust im K-Fall	IO-Performance-Auswirkungen
synchron	kein Verlust	jede Schreib-IO wird verlängert
semi-synchron	max. eine Schreib-IO pro Volume	nur schreibintensive Phasen werden verlängert
asynchron	alle Schreib-IOs seit dem letzten Kon-sistenzstand	nur minimale Auswirkungen

2.8.2 Domino-Modus

Der Domino-Modus (oder Domino-Effekt) ist zunächst ein Symmetrix-Attribut, das für ein SRDF-Paar eingestellt werden kann. Es stellt sicher, dass die Daten auf Source Unit und Target Unit stets synchron sind. Wenn dieses Attribut für ein SRDF-Paar aktiviert ist, wird das die Source Unit enthaltende Symmetrix System bei Nichtverfügbarkeit einer der beiden gespiegelten Units sowie bei einem Verbindungsfehler zwischen den beiden über SRDF verbundenen Symmetrix Systemen immer die Source Unit auf „disabled“ setzen (falls noch möglich) und jede IO auf die Source Unit mit der Fehleraussage "intervention required / unit not ready" abweisen. Die Anwendung wird sich dann mit Fehler beenden oder stehenbleiben, bis die Source Unit per Kommando reaktiviert wird, wofür zuvor der Domino-Modus deaktivert werden muss. Es ist also, auch wenn nur die SRDF-Links oder die Target Unit ausfallen, nach deren Wiederverfügbarmachung immer ein Systemverwaltereingriff nötig, der auch die Source Units wieder verfügbar macht und ggf. den Domino-Modus neu aktiviert. Allerdings ist bei Redundanz und räumlicher Trennung der SRDF-Links der Ausfall aller Links unwahrscheinlich.

Im „normalen“ Betrieb eines SRDF-Paares, wenn der Domino-Modus nicht aktiviert ist, würden bei Nichtverfügbarkeit einer der beiden gespiegelten Units oder bei einem Verbindungsfehler zwischen den beiden über SRDF verbundenen Symmetrix Systemen IOs auf das SRDF-Paar weiterhin durchgeführt, und Schreib-IOs in der Symmetrix für späteren Transfer auf die nicht verfügbare Unit (als invalid tracks, s.u.) markiert.

Wenn eine ausgefallene SRDF-Verbindung wieder hergestellt wird oder die ausgefallene Unit wieder verfügbar wird, beginnt die automatische Resynchronisation zwischen Source und Target Units, falls sie nicht durch ein entsprechendes Symmetrix Attribut verhindert wird („Prevent automatic link recovery after all link failures = no“). Bei dieser Synchronisation wird jedoch nicht mehr die Reihenfolge der ursprünglichen Schreibaufträge eingehalten, d.h. wenn es während der Resynchronisation zu einem Ausfall der Source Units (K-Fall) kommen sollte, können die Daten auf den Target Units inkonsistent sein. Diesen (allerdings höchst unwahrscheinlichen) Fall eines Doppelfehlers mit kritischen Folgen bezeichnet man auch als „Rolling Disaster mit Linkflattern“ (Abb. 2-1). In einem solchen Fall hilft dann nur noch der Rückstieg auf die letzte Sicherung, die in einem ausgelagerten Bandarchiv oder durch Duplizierung der Sicherungen im Ausweich-RZ zur Verfügung steht. Vor diesem Fall ist man durch den Domino-Modus geschützt, da bei Ausfall der SRDF-Links keine IOs mehr zugelassen werden.

Interessant ist der Domino-Modus ebenfalls für kritische Anwendungen, bei denen der Verlust von wenigen ausgeführten Transaktionen (Buchungen, Aufträge o.ä.) teure Folgen haben kann. So ist ein Rolling Disaster Szenario denkbar, bei dem zunächst die SRDF-Verbindungen ausfallen, die Anwendungen weiter IOs auf die Source Units durchführen, und erst in einem zweiten Schritt der Produktiv-Host ausfällt, so dass nach Umschaltung auf das Standby-RZ und die dortigen Target Units der Datenbestand nicht ganz aktuell ist (Entspricht der Abb. 2-1, jedoch ohne Schritt 2). Dieses Szenario ist leichter vorstellbar, als das vorher beschriebene Szenario mit Linkflattern, hat aber dafür nicht die höchst unangenehme Konsequenz der Dateninkonsistenz, sondern „nur“ die mögliche Konsequenz des Verlusts der zuletzt geschriebenen Daten.

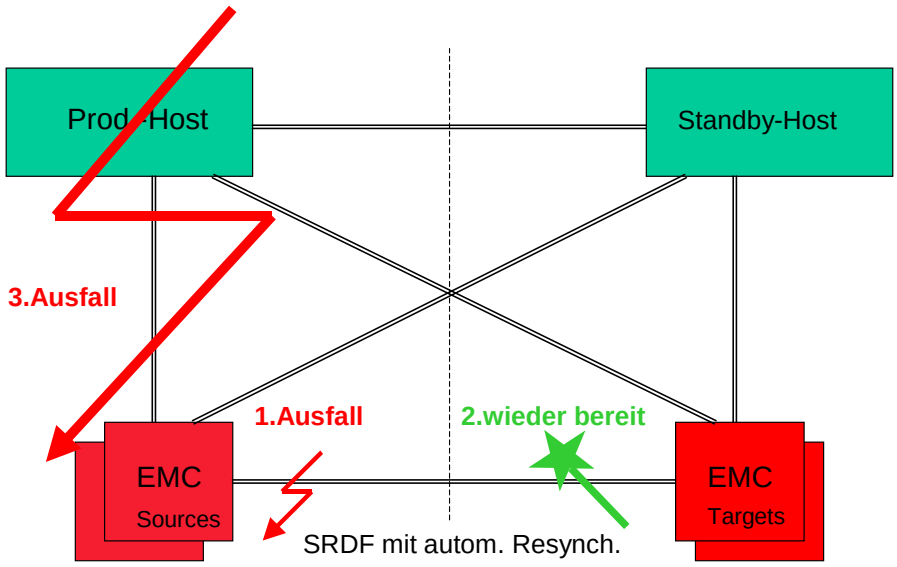


Abb. 2-1 Rolling Disaster mit „Linkflattern“

Beim Domino-Modus bewirkt allerdings schon der Ausfall der SRDF-Links sowie der Ausfall einer Target Unit den Ausfall der Source Unit und somit eine Anwendungsunterbrechung (s. hierzu auch Kap. 4.2). Dadurch ist die Hochverfügbarkeit wiederum verringert, und deshalb gilt bei der Entscheidung für oder gegen den Domino-Modus:

- Ist die höchste Verfügbarkeit der Anwendungen oberstes Gebot, dem zuliebe geringe Risiken beim Katastrophenschutz eingegangen werden können, so wird man eine KS-Konfiguration ohne Domino-Modus wählen.
- Wenn das Risiko eines Datenverlusts in (höchst unwahrscheinlichen) speziellen Katastrophenszenarien unter gar keinen Umständen eingegangen werden darf, dann muss man eine Anwendungsunterbrechung bei Einzelausfällen im Produktions-RZ, Standby-RZ oder auf dem Verbindungsweg zwischen den beiden Standorten akzeptieren und eine KS-Konfiguration mit Domino-Modus wählen.

Der Domino-Modus kann im BS2000 über Kommandos des Subsystem SHC-OSD für einzelne SRDF-Paare eingeschaltet werden.

Der Domino-Modus lässt sich verallgemeinern auf Konsistenzgruppen von Datenträgern, die über reine SRDF-Konfigurationen hinausgehen, siehe hierzu den Abschnitt 2.11. Im Abschnitt 4.2 ist insbesondere auch vom Domino-Modus für DAB-Caches in einer Dual-GS-Konfiguration die Rede.

2.8.3 Invalid Tracks

Wenn (bei ausgeschaltetem Domino-Modus) eine Target Unit ausfällt oder wenn alle Remote Links ausfallen, so markiert die lokale Symmetrix alle neuen Schreibdaten (und die ggf. noch nicht auf die Target Unit durchgeschriebenen Daten) als Invalid Tracks. Wenn der Defekt wieder behoben ist, so wird abhängig von einer speziellen Symmetrix-Einstellung (prevent automatic link recovery after all link failures = no) die Symmetrix eine automatische Resynchronisierung aller als Invalid Tracks markierten Spuren durchführen. Bei bestehender Verbindung zur remote Symmerix wird auch diese über die Anzahl der invalid tracks informiert.

Eine interne volume-spezifische Einstellung des Symmetrix Subsystems, genannt „Invalid Tracks Attribut“, verhindert, dass mit einer Target Unit alleine weitergearbeitet werden kann, wenn dieses Volume nicht synchronisiert ist. Diese Einstellung kann mit einem SHC-Parameter außer Kraft gesetzt werden, sie verhindert aber, dass invalid tracks beim Failover unbemerkt bleiben. Wenn, wie auch später in Kapitel 4 empfohlen, SRDF-Verbindungen mehrfach vorhanden sind und räumlich getrennt verlaufen, sollte der Ausfall sämtlicher Links ausschließbar sein und somit keine invalid tracks auf den Target Units auftreten. Bei asynchronen Konfigurationen muss man im K-Fall allerdings damit rechnen (s.a. Kap. 4.4).

2.8.4 SRDF-Einstellungen beim synchronen Katastrophenschutz

Es wird für alle Platten (bis auf eventuell vorhandene BCVs, siehe 2.9) in Katastrophenschutz-Konfigurationen RAID1 vorausgesetzt, so dass der Fall des Schreibfehlers auf einer einzelnen Platte (z.B. Platte mit den Logging-Daten) nicht vorkommt. Nur die Remote Link Verbindungen oder ganze Symmetrix Subsysteme müssen als Ausfalleinheiten einer

Symmetrix-Konfiguration betrachtet werden. Alle involvierten Platten müssen RAID1-gespiegelt sein, um (lokale) Hochverfügbarkeit bieten zu können.

Logisch abhängige Daten sollten nicht über verschiedene Symmetrix Subsysteme verstreut werden. Sonst sind Rolling Disaster Fälle denkbar, bei denen im Standby-RZ z.B. das Logging einer Transaktion geschrieben wurde, die Transaktion selber aber nicht (s. auch Abschnitt 4.6).
Solange der Verlust von Daten durch die in Kap. 2.8.2 beschriebenen Szenarien im K-Fall tolerierbar ist, wird empfohlen, den Domino-Modus nicht einzuschalten, um die Hochverfügbarkeit der Konfiguration nicht zu beeinträchtigen.
Beim Konfigurieren der Symmetrix-Steuerungen ist darauf zu achten, dass vom EMC-Techiker

- die automatische Resynchronisation für die Steuerungen eingeschaltet worden ist (Einstellung „Prevent automatic link recovery after all link failures = no“),
- das Invalid Tracks Attribut für alle SRDF-Platten eingeschaltet worden ist.

2.9 Zur Symmetrix-Funktion TimeFinder/Mirror

Die Funktion TimeFinder/Mirror bietet die Möglichkeit, zu einer Symmetrixplatte eine Spiegelplatte von gleichem Typ und gleicher Kapazität einzurichten, die während des Produktivbetriebs hinzugeschaltet und abgetrennt zu Test- oder Sicherungszwecken genutzt werden kann, ohne dass die Anwendungen hierfür beendet werden müssen (Point-In-Time-Copy). Die Originalplatte wird dann als Normal Unit und die Spiegelplatte als Additional Mirror Unit oder auch **B**usiness **C**ontinuance **V**olume (BCV) bezeichnet. Die Funktion ist ausführlich im SHC-OSD-Handbuch [1] beschrieben.
Ab der Version 3.0 des SHC-OSD ist es weiterhin möglich, pro Platte mehr als ein BCV alternierend in Betrieb zu nehmen („Multi BCVs“), deren Abgleich mit dem Original auf der Technik einer Differenzsicherung beruht. Eines der BCVs kann dann stets als Spiegel in Betrieb sein, während auf dem anderen beispielsweise ein konsistenter, eingefrorener Datenbestand vorliegt. Wird der aktuelle Spiegel weggeschaltet und das zweite BCV hinzugenommen, werden nur zwischenzeitlich geänderte Datenblöcke auf das zweite BCV kopiert. Da man sowohl SRDF Source Units als auch SRDF Target Units mit einem BCV-Spiegel versehen kann, ist diese Funktion auch für den (asynchronen) Katastrophenschutz nutzbar. In den Abschnitten 4.4 und 4.5 wird hierauf Bezug genommen.

2.10 Globalspeicher (GS)

Der Globalspeicher (oder Global Storage, kurz GS genannt) ist ein auf Halbleiterbasis arbeitender schneller Erweiterungsspeicher und kann an BS2000-Anlagen der S-Linie angeschlossen werden. Der GS besteht aus 1 oder 2 unabhängigen Hardware-Einheiten (GS-Units), die jeweils eine eigene Speicheransteuerung und Stromversorgung besitzen. Wenn 2 GS-Units vorhanden sind, können diese parallel betrieben werden (Dual-Modus), wobei dieselben Daten bei einer Schreiboperation auf beide GS-Units geschrieben werden. Hiermit kann man dem Ausfall einer kompletten GS-Unit vorbeugen. Eine GS-Unit wird üblicherweise auch mit einer Batterie ausgestattet, um die Daten auch bei Stromausfall aufrecht zu erhalten.

Der GS kann einerseits zur Emulation von Festplatten genutzt werden (sogenannte GS-Volumes, eine Art von RAM-Disks), die um ein Vielfaches schneller arbeiten als herkömmliche Platten, oder aber als Cache-Medium für herkömmliche BS2000-Platten. Bei einem Systemausfall bleiben die Cache-Daten im GS erhalten und werden bei Neustart des Systems rekonstruiert, so dass ein dualer GS (mit Batterie) uneinge-schränkt auch zum sicheren Schreib-Caching genutzt werden kann. Für unsere Betrachtungen ist die GS-Nutzung als Cache-Medium insofern wichtig, da die Cache-Bereiche kritischer Dateien mit in den Katastrophenschutz einbezogen werden müssen (Kap. 4.2). Genaue Informationen zur Nutzung von GS als Cache sind im DAB-Handbuch [2] zu finden.

2.11 Konsistenz-Gruppen

Der Begriff Konsistenz-Gruppen leitet sich vom speziellen EMC-Feature „Consistency Groups“ oder auch von einer gleichnamigen Funktion innerhalb des IBM-Features „XRC“ (XRC ist eine Art von asynchroner entfernter Spiegelung) ab und wird hier verallgemeinert, um insbesondere im Globalspeicher von DAB zwischengepufferte Dateien in ein KS-Konzept mitaufnehmen zu können.

Eine Konsistenzgruppe wird immer in einer Umgebung definiert, bei der als Katastrophenvorsorge eine Spiegelung von Daten zum Zwecke der Datenverfügbarkeit in einem Standby-RZ eingesetzt wird.

- Wenn die abstrakte Zusammenfassung aller physikalischen und logischen Datenträger, auf die logisch abhängige Daten einer Anwendung (Standard-Beispiel Datenbank-Anwendung, Transaktion und Logging) verteilt sind, hostseitig gemeinsam so kontrolliert und gesteuert werden kann, dass Datenintegrität und Daten-Konsistenz der Anwendung auch bei beliebigen Ausfallszenarien auf den gespiegelten Datenträgern im Standby-RZ gewährleistet sind, so sprechen wir von einer **Konsistenzgruppe**.

Beispiele von Konsistenzgruppen:

- Daten einer Anwendung können über mehrere Symmetrix Source Units innerhalb eines Symmetrix Systems verteilt sein. Sind die Source Units jeweils RAID1-geschützt, so liegt eine Konsistenzgruppe vor.
- Daten einer Anwendung können über mehr als ein Symmetrix System verteilt sein. Um eine Konsistenzgruppe zu erhalten, muss entweder das EMC-Konzept der „Consistency Groups“ genutzt werden (s.u.), oder die Daten müssen so auf die Symmetrix Systeme verteilt werden, dass logisch abhängige IOs immer dieselbe Symmetrix ansprechen.
- Datenbanken setzen eigene Schreib-Caches (im Hauptspeicher, also „flüchtige“ Caches) ein, deren Synchronisation mit den Platten durch Datenbank-spezifische Algorithmen erfolgt. Wenn die Datenbank sicherstellt, dass nach einem Hostausfall die Datenbank über Before-Images konsistent rekonstruiert werden kann, so bilden Platten und Cache der Datenbank eine Konsistenzgruppe.

Beispiel für das Fehlen einer Konsistenzgruppe:
- Daten einer Anwendung können in einem Symmetrix System und einem nichtflüchtigen Schreib-Cache einer System-SW verteilt sein (DAB Schreib-Cache im GS, siehe 4.2). Eine solche Verteilung stellt i.d.R. keine Konsistenzgruppe dar, weil die beiden Subsysteme Symmetrix und Caching-SW nicht die dafür erforderliche intime Kommunikation pflegen.
- Ebenso können Daten einer Anwendung in einem Symmetrix System und einem GS-Volume (s.2.10) verteilt sein. Eine solche Verteilung stellt i.d.R. keine Konsistenzgruppe dar, weil die beiden Subsysteme Symmetrix und die GSVOL-SW ebensowenig miteinander kommunizieren.

Bei EMC (Symmetrix) und IBM (ESS) bezieht sich die „Consistency Group“ ebenfalls auf Umgebungen, bei denen eine Spiegelung von Daten (SRDF, XRC) zum Zwecke der Datenverfügbarkeit in einem Standby-RZ auch in einem Katastrophenfall eingesetzt wird:

- EMC definiert Konsistenz-Gruppen als Gruppen von SRDF Devices, deren Zusammenfassung der kontrollierten Daten-Integrität und Daten-Konsistenz einer über mehrere SRDF Units verteilten Datenbank dient. Die SRDF Source Units einer Konsistenz-Gruppe können dabei über mehrere Symmetrix Systeme verteilt sein. Realisiert ist eine Konsistenz-Gruppe dadurch, dass bei Auftreten eines nicht behebbaren Fehlers für eine IO auf ein Source/Target-Paar der Konsistenz-Gruppe (z.B. keine Kommunikation mit der Target Unit möglich) ein Konsistenzprotokoll dafür sorgt, dass für alle Folge-IOs auf Devices der Konsistenz-Gruppe die SRDF-Spiegelung ausgesetzt wird. Damit wird verhindert, dass der entfernte Datenbestand auf den Target Units der Konsistenzgruppe inkonsistent wird.
- IBM definiert Konsistenz-Gruppen als Gruppen von Datensätzen von XRC Dateien, deren Übertragung auf die Target Units der entfernten ESS Systeme garantiert in der Original-Reihenfolge der Anwendungen stattfindet. Die Einhaltung der Reihenfolge der Schreib-IOs ist von fundamentaler Bedeutung für abhängige IOs auf verschiedene devices (Standard-Beispiel wie oben: Datenbank-Anwendung, Transaktion und Logging).

2.12 HIPLEX MSCF

HIPLEX (**H**ighly **I**ntegrated System Com**plex**, beschrieben in [4]) ist das Konzept von Fujitsu zur Realisierung eines Verfügbarkeits- und Leistungsverbundes von mehreren BS2000/OSD Business Servern. Das Softwareprodukt HIPLEX MSCF (MSCF = **M**ultiple **S**ystem **C**ontrol **F**acility) stellt die für einen Leistungs- und Verfügbarkeitsverbund (auch MSCF-Verbund genannt) erforderliche Infrastruktur sowie Basisfunktionen für verteilte Anwendungen bereit.

Grundlegend in einem MSCF-Verbund ist die Kommunikation der beteiligten Rechner auf der Basis von BCAM-Transportverbindungen. Zwischen den Rechnern werden Aufträge zur Ausführung von Funktionen und Kontrollnachrichten zur Überwachung des Verbunds ausgetauscht.

Im MSCF-Verbund ist der Shared Pubset-Verbund der wichtigste Verbundtyp. Zusätzlich zur Kommunikationsverbindung haben bei diesem Verbundtyp alle Rechner des Verbunds Zugriff auf gemeinsame Platten, das Shared Pubset. Mittels zweier unabhängiger Datenwege, also mithilfe der gemeinsamen Platten einerseits und Kommunikationspfaden andererseits, überwachen sich die Rechner des Verbunds gegenseitig, wobei durch die Redundanz der Überwachungspfade ein Rechnerausfall zuverlässig erkannt werden kann. Beim Ausfall eines Rechners oder eines Kommunikationspfades gewährleisten geeignete Rekonfigurierungsmaßnahmen die weitere Funktionsfähigkeit des Shared-Pubset-Verbunds.

Der XCS-Verbund (Cross Coupled System) ist eine Erweiterung des Shared-Pubset-Verbunds. Rechner-übergreifende Synchronisationsmechanismen erlauben die Verwaltung global verfügbarer Betriebsmittel und ermöglichen den Betrieb von verteilten Anwendungen mit Datenzugriff auf die gemeinsamen Datenträger.

Die Mechanismen zur Erkennung eines Rechnerausfalls sind integraler Bestandteil sowohl des Shared-Pubset-Verbunds und des XCS-Verbunds als auch des CCS-Verbunds ((Closely Coupled System) ohne Shared Pubset. Sie erlauben die Realisierung von Standby-Konfigurationen, mit deren Hilfe sich Ausfallzeiten von Anwendungen minimieren lassen. Die Realisierung derartiger Konfigurationen wird durch ein weiteres Mitglied der HIPLEX-Produktfamilie, HIPLEX **A**vailability **F**acility (HIPLEX AF), unterstützt.

Zusätzlich zur Überwachung über die BCAM-Verbindungen überwachen sich die Teilnehmer eines Shared-Pubset-Verbunds über Shared Pubsets. Dazu wird beim Importieren eines Pubsets im Shared-Modus auf dem Master-Rechner vom DVS die Watch-Dog-Datei automatisch eingerichtet. Jeder Sharer (sowohl Master- als auch Slave-Rechner) schreibt in die Watch-Dog-Datei periodisch einen inkrementierten Zähler als Lebendmeldung und liest die Lebendmeldungen der anderen Sharer (Plattenprotokoll). Ein potenzieller Ausfall eines Sharers wird daran erkannt, dass von ihm keine neue Lebendmeldung mehr geschrieben, d.h. sein Zähler nicht mehr inkrementiert wird.

2.13 HIPLEX AF

Das Softwareprodukt HIPLEX **AF** (**A**vailability **F**acility) erhöht die Verfügbarkeit von Anwendungen bei einem Anwendungs- oder Systemausfall in einem Verbund mehrerer BS2000/OSD Business Server.

Der Einsatz von HIPLEX AF beruht auf dem Prinzip der Redundanz. Anstelle **eines** Systems, welches den Lastbedarf aller Anwendungen abdeckt, sind **mehrere** Systeme installiert, die sich im normalen Betrieb die Last teilen. Fällt eines dieser Systeme völlig oder teilweise aus, so können intakte Systeme die wichtigsten Anwendungen übernehmen, wenn auch möglicherweise mit reduzierter Performance. Voraussetzung hierfür ist, dass die umzuschaltenden Anwendungen und ihre Ressourcen entsprechend konfiguriert sind.

Im normalen Betrieb ist der Einsatz von HIPLEX AF für die Anwender der überwachten Systeme nicht merklich. Bei Systemausfall veranlasst HIPLEX AF die automatische Umschaltung von laufenden Anwendungen des Work-Systems auf ein Standby-System. Dieses automatische Umschalten erfolgt sehr schnell und auch im unbedienten Betrieb.

Die Anwendungen des Work-Systems können zudem zu jedem beliebigen Zeitpunkt manuell auf ein Standby-System mit einem Kommando von HIPLEX AF umgeschaltet werden, beispielsweise vor der Wartung des Work-Systems oder bei Einführung einer neuen Software-Version.

HIPLEX AF minimiert die Ausfallzeit der Anwendungen durch folgende Faktoren:
- Die eingesparte Zeit durch die automatische Ausfallerkennung. Im unbedienten Betrieb ist dieser Zeitgewinn entscheidend für die rasche Wiederaufnahme des Produktivbetriebs.
- Die eingesparte Zeit für den Neustart des Work-Systems.
- Die eingesparte Zeit für eine eventuelle Hardware-Reparatur.
- Die Zuverlässigkeit der Umschaltung, die durch ausgetestete Verfahren gewährleistet wird.

HIPLEX AF nutzt PROP-XT-Administrationsprozeduren, die auf den Work-und Standby-Systemen des Shared Pubset-Verbundes gestartet werden. Die Administrationsprozeduren von HIPLEX AF kommunizieren untereinander mittels Jobvariablen auf dem Shared Pubset.

Voraussetzung für die optimale Nutzung von HIPLEX AF für Katastrophenschutz-Konzepte ist ein Shared-Pubset-Verbund und der Einsatz von HIPLEX MSCF, und damit insbesondere eine Konfiguration der in Abb. 2-2 gezeigten Art:

Prinzipbild einer optimalen „HIPLEX AF fähigen“ KS-Konfiguration

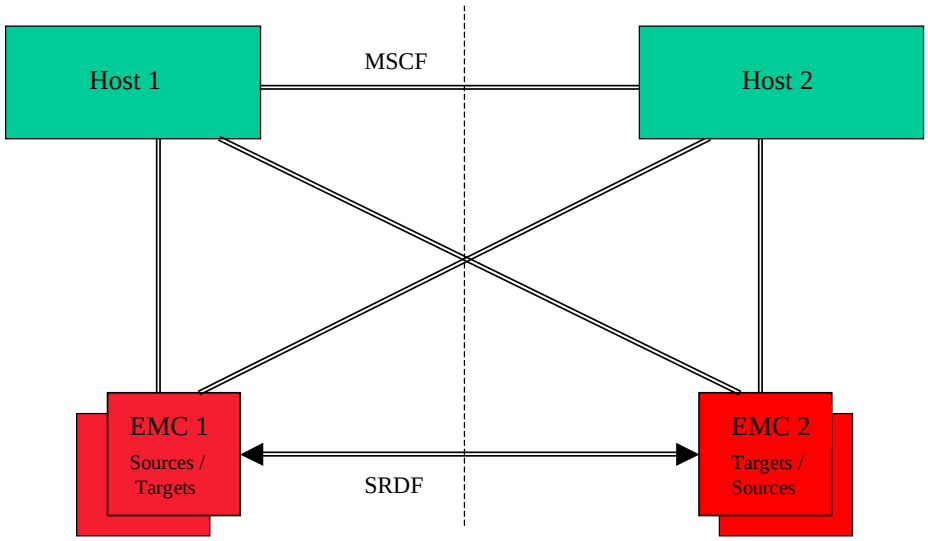


Abb -2.2 Automatisierbare HIPLEX AF fähige KS-Konfiguration

Eine solche Konfiguration wird wegen der „Kreuzverkabelung“ zwischen Hosts und Plattenspeicher-Systemen auch X-Link-Konfiguration genannt.

Mit HIPLEX AF wird zwischen den zwei Hosts ein Hochverfügbarkeitsverbund geschaffen, in dem die Systeme sich gegenseitig und auch die Anwendungen überwachen. Es werden eine oder mehrere Anwendungen und deren Hard- und Software-Ressourcen als eine Umschalteinheit („Switch-Unit“) definiert. Die Überwachung der Anwendungen und ihrer Ressourcen durch HIPLEX AF wird als „Monitoring“ der Switch-Unit bezeichnet. Nach vom Kunden festgelegten Kriterien wird die Switch-Unit bei Systemausfall oder Ausfall wichtiger Ressourcen auf den Standby-Host umgeschaltet („Switchover“).

In HIPLEX AF ist es weiterhin möglich, auch den Ausfall eines Symmetrix Systems zu erkennen und automatisch die Target Units zu aktivieren und mit diesen die Anwendung neu anzustarten (Failover). Bei Einsatz von HIPLEX AF zum automatischen KS wird die Entscheidung darüber, ob ein Failover durchzuführen ist, von der Überwachungssoftware gefällt, und es kann auch im unbedienten Betrieb ein Failover erfolgen.

Möglich ist es auch, mehrere Anwendungen, die auf Work- und Standby-Host laufen, als eigene Umschalteinheiten (Switch-Units) zu definieren, die dann getrennt voneinander umgeschaltet werden können. In diesen Switch-Unit-Definitionen können dann Listen von einzelnen Target Units angegeben werden, die bei einer Umschaltung halb- oder vollautomatisch freigeschaltet werden.

2.14 Storage Host Component im BS2000 (SHC-OSD)

SHC-OSD ist die Steuerungssoftware für Symmetrix Systeme im BS2000. Zusammen mit der zugehörigen Symmetrix-Firmware (Microcode) kann man damit eine ins Betriebssystem BS2000 integrierte Steuerung der Symmetrix-Funktionen SRDF und TimeFinder durchführen, oder die Konfigurationsdaten der Symmetrix ermitteln. Das Produkt ist im Handbuch [1] ausführlich beschrieben.

Das Einrichten von SRDF-Units sowie von BCV Additional Mirror Units ist Aufgabe der EMC-Techniker.

Ab der Version SHC-OSD V4.0 steht ein Tool zur Verfügung, mit dem die für Failover und Failback erforderlichen Symmetrix-Aktionen sowohl manuell per Kommando (resp. Prozedur) als auch automatisiert unter Kontrolle von HIPLEX AF realisiert werden können. Für ein SRDF-Plattenpaar kann der Domino-Effekt mit einem SHC-OSD-Kommando ein- resp. ausgeschaltet werden. Darüber hinaus bietet SHC-OSD Überwachungsfunktionen, die Zustandsänderungen der Konfiguration der Symmetrix-Steuerungen, den Status der Geräte und den Status des Remote-Copy-Betriebes anzeigen. Werden Zustandsänderungen erkannt, wird durch SHC-OSD eine auffällige, zu beantwortende Meldung auf der Konsole ausgegeben.

2.15 Dual Recording By Volume (DRV)

DRV (Dual Recording by Volume) ist ein Software-gesteuertes Aufzeichnungsverfahren, mit dem die Daten auf zwei Platten doppelt geführt werden können. Zur Unterscheidung wird das einfache Aufzeichnungsverfahren SRV (Single Recording by Volume) genannt.

DRV wird im E/A-System des BS2000 realisiert und muss weder vom Data Management System noch vom Anwenderprogramm zur Kenntnis genommen werden. Der DRV-Betrieb wird über eine Reihe von Kommandos vom Operator bzw. Systemverwalter eingeleitet, gesteuert, überwacht und beendet.

Der DRV-Modus, bei dem die Daten doppelt geführt werden, heißt Dual-Modus. Er erhöht die Verfügbarkeit der auf den Platten gespeicherten Daten. Jeder Schreibauftrag des Data Management System wird auf beiden Platten ausgeführt, und jeder Leseauftrag wird von der Platte mit der kürzesten IO-Warteschlange oder alternativ von einer fest ausgewählten Platte abgewickelt.

Fällt eine Platte aus, kann auf Mono-Modus gewechselt werden. Ein Eingriff des Operators ist dafür nicht notwendig. Der Operator oder Systemverwalter kann das fehlerhafte Laufwerk durch ein anderes Laufwerk vom gleichen Typ ohne Unterbrechung der Anwendungen austauschen. Der Mono-Modus unterscheidet sich von SRV dadurch, dass während des Betriebs durch Zuschalten einer Platte mit identischer Volume Serial Number der Dual-Modus wieder aufgenommen werden kann. Der Übergang von Mono- auf Dual-Modus heißt Rekonstruktion. Die Daten werden auf die hinzugenommene Platte kopiert, wobei Ein-/Ausgaben der Benutzer gleichzeitig bearbeitet werden können.

DRV unterstützt Pubsets und Privatplatten aller Gerätetypen mit der einzigen Einschränkung, dass Shared Pubsets nicht für DRV verwendet werden können.

Durch die räumliche Trennung von Original- und Spiegelplatte, die theoretisch von dem BS2000-Server auch weit entfernt sein können, ist der Datenzugriff trotz Nichtverfügbarkeit eines Plattenspeichersystems z.B. im K-Fall unterbrechungsfrei möglich. Damit steht mit dem BS2000-Subsystem DRV ein Software-Tool zur Verfügung, das ähnliche Funktionalität wie SRDF bietet und auch für einen Katastrophenschutz alternativ zu SRDF eingesetzt werden kann.

Während im K-Fall die SRDF-Targets über SHC-OSD-Kommandos zur Benutzung freigeschaltet werden müssen, können die DRV-Spiegelplatten nach UNLOCK-DISK vom Standby-System direkt genutzt werden.
Die funktional mächtigere Variante ist die SRDF-Variante.
Dennoch ist DRV für Katastrophenschutz-Konzepte eine erwägenswerte Möglichkeit, die insbesondere bei ESCON-Anschluss für Katastrophenschutz bei Konfigurationen mit geringer Entfernung zwischen Produktions- und Standby-RZ in Frage kommt.
Um das vorliegende Papier nicht zu „überfrachten“, wurden keine eigenen Katastrophenschutz-Konfigurationen mit DRV-Spiegelung statt SRDF-Spiegelung im Abschnitt 4 aufgenommen. Bei einem konkreten Projekt können sie jedoch jederzeit beschrieben und diskutiert werden.

	SRDF	DRV
Kopier-Modi	Synchron, Semi-Synchron, Asynchron	Synchron
Verlängerung einer Schreib-IO	Beim Synchronen Modus: + Signallaufzeit über den SRDF-Link + Einlagerung der Daten in den Cache der entfernten Symmetrix	Zwei IOs werden parallel durchgeführt, der zweite Interrupt bestimmt die Gesamt-IO-Dauer. Bei kleinen Entfernungen performanter als SRDF.
Verkabelung	Kopie wird über den SRDF-Link ausgeführt	Kopie wird über eine zweite ESCON-(oder FC-)Verbindung ausgeführt, Kreuzverkabelung zwingend !
Entfernung Prod.-RZ zu Standby-RZ	Auch bei großen Entfernungen möglich	Die Entfernung bei ESCON-Anschluss sollte 9 km nicht überschreiten. Große Entfernungen bei FC Anschluss möglich
Spiegelung von Shared Pubsets	Möglich	Nicht möglich
Spiegelung von Home Pubsets	Möglich	Möglich

Konsistenzgruppen	RAID1-Platten und logisch abhängige Daten nicht über mehr als ein Symmetrix System verteilen	RAID1-Platten und logisch abhängige Daten nicht über mehr als ein Symmetrix System verteilen
Benötigte Betriebsmittel	SRDF-Link	CPU-Leistung, Hauptspeicher, Kanal
Kosten	SRDF-Lizenz	Mietkosten für Subsystem DRV

Tab. 2-1 Gegenüberstellung von SRDF und DRV

3 Voraussetzungen für Katastrophenschutz im BS2000

In diesem Kapitel werden die grundlegenden Voraussetzungen für KS-Konfigurationen wie die Datenspiegelung und die Netzanbindung an Work- und Standby-Hosts beschrieben. Aufgrund der vielen möglichen Einflussfaktoren, die wir in der Einleitung erwähnt haben, wird nur ein Überblick geboten. Es werden Verfahren zur Verfügbarkeit der Online-Daten im Ausweich-RZ und Techniken zur Überbrückung weiter Entfernungen für Datenspiegelungen diskutiert und bewertet. Darüber hinaus werden organisatorische und administrative Maßnahmen angesprochen, die für ein KS-Konzept zu beachten sind.

3.1 Organisatorische Maßnahmen

Zunächst wollen wir auf organisatorische Maßnahmen hinweisen, die für jedes KS-Konzept zu treffen sind. Diese sind weitgehend abhängig vom betroffenen RZ und dessen Mitarbeiter sowie der vorhandenen Infrastruktur.

3.1.1 Notfall-Vorsorge

Zum Thema Notfall-Vorsorge gibt es nach Auffassung des Bundesamtes für Sicherheit in der Informationstechnik die folgenden Unterthemen, die vom IT-Betreiber berücksichtigt werden sollten (siehe auch IT-Grundschutzhandbuch [6]).

- Erstellung einer Übersicht über Verfügbarkeitsanforderungen
- Notfall-Definition, Notfall-Verantwortlicher
- Erstellung eines Notfall-Handbuches
- Dokumentation der Kapazitätsanforderungen der IT-Anwendungen
- Definition des eingeschränkten IT-Betriebs
- Untersuchung interner und externer Ausweichmöglichkeiten
- Regelung der Verantwortung im Notfall
- Alarmierungsplan
- Erstellung eines Wiederanlaufplans
- Erstellung eines Datensicherungsplans
- Ersatzbeschaffungsplan
- Durchführung von Notfallübungen

Durch eine automatisierte Ausfallerkennung und einen automatisierten oder halbautomatisierten Wiederanlauf der Produktiv-Anwendungen im Ausweich-RZ, wie sie das Produkt HIPLEX AF für BS2000-Systeme bietet, ist bereits das Thema Wiederanlaufplan abgedeckt, jedoch sollte zusätzlich ein Notfall-Handbuch erstellt werden, in dem außer den Zuständigkeiten auch die durch HIPLEX AF automatisierten Schritte zum Wiederanstart der Anwendungen dokumentiert sind.

3.1.2 Notfall-Handbuch

Ein Notfall-Handbuch beschreibt die manuelle Vorgehensweise in den beiden betroffenen Rechenzentren in einem Katastrophen-Fall und ist in nahezu jeder Beziehung Kunden- und Anwendungs-spezifisch.
Um einen klaren und eindeutigen Ablauf des Wiederanlaufs der Anwendungen im K-Fall zu gewährleisten, werden für die handelnden Personen Rollen definiert und abgegrenzt. Die jeweilige Rolle bestimmt die Kompetenzen und Aufgaben im Rahmen des Wiederanlaufs. Die Rollen müssen definiert und festgelegt werden.

Solche Rollen könnten z.B. sein

- K-Fall-Verantwortlicher
- K-Fall-Manager
- K-Fall-Netzadministrator
- K-Fall-Systemadministrator
- K-Fall-Anwendungsbetreuer

Beispielsweise werden eine oder mehrere Personen als K-Fall-Verantwortliche bestimmt, die befugt und kompetent sind, zu entscheiden, dass die Umstände im Work-RZ eine Umschaltung erfordern, also den K-Fall „ausrufen“. Weiterhin können K-Fall-Manager bestimmt werden, die nach dieser Entscheidung das weitere Vorgehen koordinieren und auf deren Anweisung z.B. ein K-Fall-Systemadministrator die nötigen Aktionen ausführt.

In einem Notfall-Handbuch können dann in Checklisten die Aktionen und Rückmeldungen rollenspezifisch angegeben werden. Bei der Erstellung eines Notfall-Handbuches kann Fujitsu auf Anfrage Unterstützung leisten – wie auch schon durch unsere Competence Centers geschehen.

3.1.3 Regelmäßige Notfallübungen

Auszug aus dem IT-Grundschutzhandbuch des Bundesamts für Sicherheit in der Informationstechnik
(siehe [6]):
Notfallübungen dienen der Prüfung der Wirksamkeit von Maßnahmen im Bereich der Notfallvorsorge. Einerseits wird durch eine Notfallübung der effektive und reibungslose Ablauf eines Notfall-Plans erprobt und andererseits werden bisher unerkannte Mängel aufgedeckt. Typische Übungen sind:

- die Durchführung einer Alarmierung
- Durchführung von Brandschutzübungen
- Funktionstests von Stromaggregaten
- Wiederanlauf nach Ausfall einer ausgewählter IT-Komponenten und
- Wiedereinspielen von Datensicherungen

Die Ergebnisse einer Notfallübung sind zu dokumentieren.

Notfallübungen sind regelmäßig zu wiederholen. Da diese Übungen den normalen Betriebsablauf stören können, sollte die Häufigkeit an der Gefährdungslage orientiert sein, jedoch sollten die entsprechenden Notfallübungen zumindest einmal jährlich stattfinden. Soweit erforderlich sind Schulungsmaßnahmen der Mitarbeiter durchzuführen (Erste Hilfe, Brandbekämpfung etc.). Vor Durchführung einer Notfallübung ist das Einverständnis der Behörden- bzw. Unternehmensleitung einzuholen.

Ergänzende Kontrollfragen:
- Werden die Notfallübungen regelmäßig wiederholt?
- Führen aufgedeckte Mängel zu einer Überarbeitung der Notfall-Pläne / des Notfall-Handbuchs?

3.2 Aufbau eines Standby-RZ

Ein Standby-RZ muss so ausgestattet werden, dass es die Summe der relevanten Produktivanwendungen (des Work-RZs) zum Ablauf bringen kann. Analog dazu müssen in einer symmetrischen KS-Konfiguration beide RZs die Summe aller relevanten Produktivanwendungen beider RZs zum Ablauf bringen können. Gemeint sind insbesondere die RPF-Leistung der Hosts, der Speicherausbau, der Kanaldurchsatz und die Plattenkapazität. Die RPF-Leistung kann kostensparend als „Capacity on Demand“ mit „hot-extra-CPUs“ bereitgestellt werden. Hierfür muss der Hardware-Ressourcenbedarf aller betroffenen Anwendungen ermittelt werden. Falls mehrere Anwendungen umschaltbar gemacht werden sollen, müssen genügend LAN-Anschlüsse bzw HNCs (Highspeed Net Connect, Netzanschluss für BS2000/OSD-Systeme) vorhanden sein, damit man die nötige Anzahl von IP-Adressen bereitstellen kann. Ebenso muss auch die Konfiguration der Netze an beiden Standorten erweitert werden, damit der reibungslose Betrieb auch im K-Fall mit zusätzlichen Anwendern aufrecht erhalten werden kann. Ebenso müssen ggf. zusätzliche Bandgeräte und Drucker bereitgestellt werden, um in einem K-Fall die erforderliche Bandgeräte- und Drucker-Leistung auch im Ausweich-RZ zur Verfügung zu haben.

Zusammengefasst kann man sagen, dass zwei etwa gleich ausgerüstete (spiegelbildliche) RZs vorhanden sein müssen.

3.3 Vorbereitung Datennetz

Es muss eine Netzanbindung der Anwender an den jeweiligen Standby-Host vorhanden sein. Wünschenswert ist natürlich eine Konfiguration, die es mit wenigen oder gar keinen Eingriffen an Netzkomponenten (Switches, Router, usw.) ermöglicht, alle betroffenen Anwender in einem K-Fall auf den Standby-Host umzuleiten. Ebenso wünschenswert wird es bei vielen Anwendungen sein, dass diese Umleitung ohne Eingriffe bei den Clients vor sich geht, da diese unter Umständen größere Aufwände oder gar Neuinstallationen bedeuten können. In den folgenden zwei Kapiteln werden hierfür Möglichkeiten skizziert, außerdem wird das Thema Firewalls angesprochen. Die vielfältigen Möglichkeiten der Netztopologien und die dafür nötige Hardware werden in diesem Dokument jedoch nicht erörtert, eine Lösung für den KS wird i.d.R. von dem bereits vorhandenen Equipment abgeleitet werden können.

3.3.1 Umleitung der Netzanbindung

Die Netzanbindung beider Systeme (Hosts oder Menge von VMs) sollte idealerweise so ausgelegt sein, dass den Nutzern, die sich i.d.R. an einem dritten oder an unterschiedlichen Standorten befinden, der Zugang zu der Anwendung ermöglicht wird, unabhängig davon, auf welchem der zwei Hosts diese gerade läuft. Weiterhin muss dafür gesorgt werden, dass die Netzkomponenten wie Router, Switches, HUBs, HNCs in beiden RZs so dimensioniert sind, dass sie die Zugriffe von allen Nutzern beider Hosts auch in einem der RZs (im K-Fall) verarbeiten können.

Für die Umschaltung der Netzanbindung gibt es unterschiedliche Vorgehensweisen wobei man je nach Art der (bereits vorhandenen) Netztopologie entscheiden muss, welche die für den Kunden am besten geeignete ist. Wir wollen hier im folgenden zwei grundsätzliche Möglichkeiten in vereinfachter Form beschreiben:

Nutzung von virtuellen Hosts und dynamischem Routing
Wenn eine Anwendung auf ein anderes System umgeschaltet wird, ändert sich, physikalisch gesehen, der „Stecker“, über den die Anwendung mit ihrer Umwelt im Netz kommuniziert. Diese Änderung können Sie vor den Benutzern verbergen. Dazu können (in BS2000-Umgebungen) zusätzlich zum realen Host **virtuelle Hosts** definiert und die Anwendungen mit den virtuellen Hosts gekoppelt werden. Um Anwendungen auch nach dem Neustart auf einem Standby-System mit der gleichen Netzadresse adressieren zu können, muss im Netz der Hostname unverändert bleiben. Das kann erreicht werden, indem die Anwendung, deren Netzadresse unverändert, also transparent für die User, bleiben soll, auf einem virtuellen Host eröffnet wird. Dieser virtuelle Host muss auch auf dem Standby-System der KS-Konfiguration als virtueller Host generiert werden. Auch in UTM-Generierungen wird der Name des virtuellen Hosts eingetragen. Aus dem Netz heraus darf nur ein einziges System mit dem Hostnamen dieses virtuellen Hosts sichtbar sein, d.h. der virtuelle Host darf nur an einem realen Host aktiv sein. Die Nutzung von virtuellen Hosts stellt die transparenteste Lösung für die betroffenen Anwender dar. Voraussetzung hierfür ist, dass die IP-Adressen von Work- und Standby-Host entweder im gleichen (physikalischen) LAN-Segment liegen oder, dass für die virtuelle(n) Hostadresse(n) in den Routern des Standby-RZ entsprechende Routing-Einträge vorgenommen werden. Weitere Erläuterungen zum Thema virtuelle Hosts sind im Handbuch für HIPLEX AF [3] zu finden.

Die Abb. 3-1zeigt ein vereinfachtes Beispiel einer Konfiguration, wie sie innerhalb eines Campus vorliegen könnte: Hier liegen die 2 Hosts im selben logischen Subnetz 139.20.10.x, wobei die LANs beider RZs beispielsweise über einen ATM-Backbone miteinander verbunden sind. Die beiden Hosts könnten auch in unterschiedlichen Subnetzen liegen und über unterschiedliche Router erreichbar sein. Im K-Fall wird der virtuelle Host 139.20.10.5 am Host B gestartet und vom BCAM werden die entsprechenden ARP-Pakete („Adress-Resolution-Protokoll“) gesendet, so dass der virtuelle Host danach in diesem LAN-Segment wieder bekannt ist, allerdings an einem Netzanschluss des realen Hosts B. Bei dieser Art der Konfiguration wäre die Netzumschaltung innerhalb kürzester Zeit (wenige Millisekunden) durchgeführt, ohne dass Änderungen in Routern, Partner-Servern oder bei den Anwendern (Clients) vorgenommen werden müssen. Für ein K-Fall Konzept empfehlen wir, wenn möglich, die betroffenen Hosts in ein Subnetz zu legen. Wenn zur Datenspiegelung, wie in Kap. 3.4 empfohlen, eine WDM-Verbindung genutzt wird, kann diese auch gleichzeitig zur (breitbandigen) Kopplung der beiden physikalischen LANs verwendet werden.

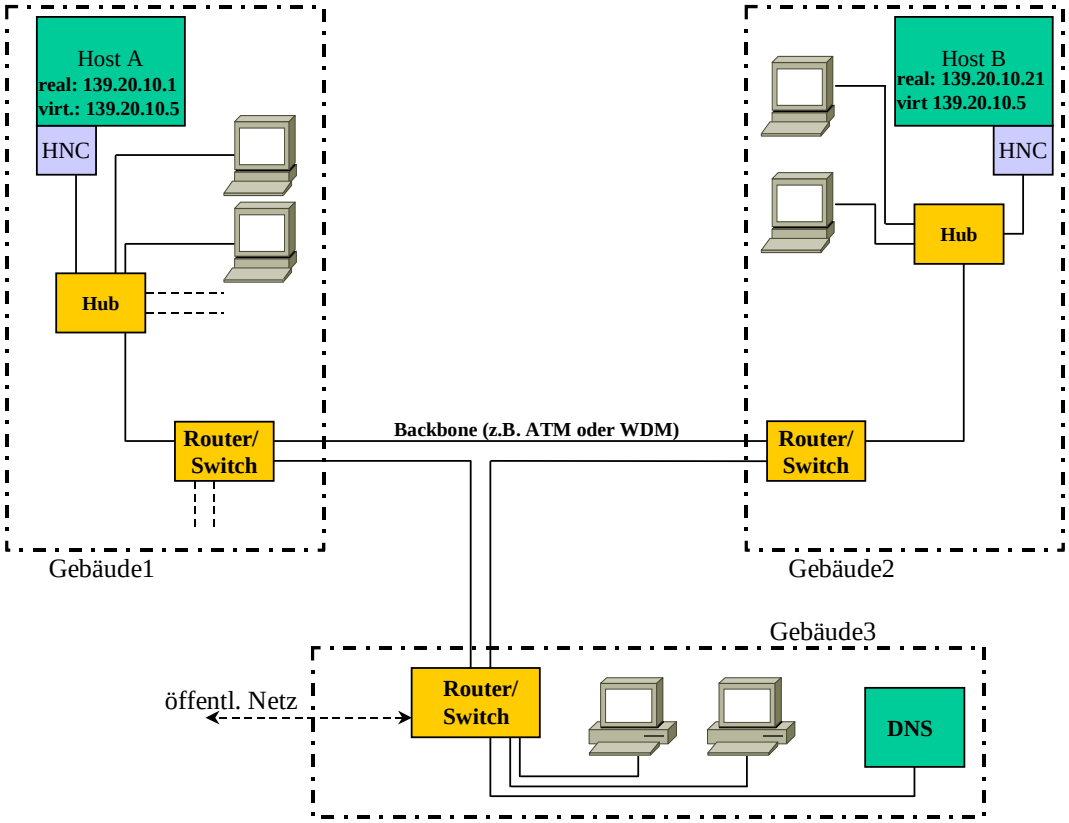


Abb. 3-1 Netzanbindung mit virtuellem Host

Falls beide Gebäude in unterschiedlichen logischen Subnetzen liegen, müssen zur Vorbereitung die virtuellen Hosts in den Tabellen aller beteiligten Router eingetragen werden, damit die Zugriffe auf auch in das Gebäude2 geroutet werden. Hierbei werden allerdings die beiden (logischen) Subnetze miteinander „vermischt". Falls mehr als die in der Abb. gezeigten 3 Router im Spiel sind, sollten diese mit einem dynamischen Routing-Protokoll arbeiten, wie z.B. OSPF („Open Shortest Path First", RFC 2328 usw.) oder IGRP („Interior Gateway Routing Protocol"), welche eine schnelle Konvergenz des Routings erreichen. Das Umkonfigurieren des Routings erst im K-Fall ist sicher nicht akzeptabel.

Eine Möglichkeit, die Transparenz der Netzadressen für die Anwender herzustellen und die ggf. unterschiedlichen LAN-Segmente logisch voneinander zu trennen, ist der Einsatz von virtuellen LANs (VLAN nach IEEE 8021.Q). Hierfür werden spezielle Layer3-Switche benötigt, wie z.B. CISCO 4908G-L3.

Nutzung von DNS
Eine andere Möglichkeit ist das Arbeiten mit DNS (Dynamic Name Server). Wenn die Nutzer ihre Anwendung (bzw. den Host) über DNS-Namen adressieren, müssen im K-Fall die Einträge im DNS-Server des Anwender-Netzes die IP-Adressen der Zielsysteme geändert und nötigenfalls ein Zonentransfer ausgeführt werden. Die für den K-Fall benötigten Adressen bzw. Subnetze müssen vorher festgelegt werden. Somit ändert sich auf Ebene der Netzadressierung nichts. Wenn jedoch viele DNS-Zonen beteiligt sind, kann dieses Vorgehen auch unübersichtlich werden. Das Verfahren stellt nur dann eine Lösung dar, wenn alle umzuschaltenden Anwendungen DNS nutzen.

Die Umschaltung durch Änderung der Einträge der IP-Adressen der Zielsysteme im DNS-Server des Anwender-Netzes wird nicht im Rahmen von automatischen Umschaltungen durch HIPLEX AF angeboten.

3.3.2 Firewalls

Falls zwei RZs in die KS-Konfiguration einbezogen sind, werden die Hosts sich auch hinter unterschiedlichen Firewallsystemen befinden. Die Security-Policies müssen zwischen beiden betroffenen RZs abgeglichen werden. Falls z.B. bestimmte Dienste im LAN-Segment des Standby-Host gesperrt sind, die die Anwender des Produktivsystems am Work-Host benötigen, wird man diese Restriktionen aufheben oder anders gestalten müssen. Wird mit „Access Control Listen" (ACLs) gearbeitet, müssen diese erweitert werden. Da es für die Konfiguration von Firewalls beliebig viele Möglichkeiten und Parameter gibt, soll dieses Thema hier nicht weiter vertieft werden. Wir wollen nur darauf hinweisen, dass das Thema Firewalls frühzeitig in eine KS-Planung mit einbezogen werden muss.

3.4 Verfahren zur Verfügbarkeit der Online-Daten im Standby-RZ

Ein wesentlicher Aspekt beim Katastrophenschutz ist die Bereitstellung der Daten der zu verlagernden Anwendungen im jeweiligen Ausweich-RZ. Denn im K-Fall muss davon ausgegangen werden, dass auch die aktuellen Daten sämtlicher Anwendungen, die sich auf den Online-Datenträgern, den Plattenspeicher-Systemen, befinden, nicht mehr verfügbar sind. Im schlimmsten Fall sind die Plattenspeicher-Systeme zerstört und damit die dort befindlichen Daten unwiederbringlich verloren.

In vielen Dokumenten verschiedener Hersteller (z.B. im Redbook „IBM System Storage Business Continuity: Part 1 Planning Guide“ von IBM [8]) wird die K-Fall-Vorsorge in 7 unterschiedliche Kategorien (tiers) spezifiziert, die aufsteigend nach Komfort und Kosten wie folgt geordnet sind:

- Tier 0: keine K-Vorsorge
- Tier 1: regelmäßiges Auslagern von Backup-Sicherungen PTAM („Pickup Truck Access Method“). Backup-Plattform wird erst im K-Fall etabliert
- Tier 2: regelmäßiges Auslagern von Backup-Sicherungen. Backup-Plattform existiert
- Tier 3: Ausgewählte Dateien werden täglich auf elektronischem Weg auf die Backup-Plattform gebracht
- Tier 4: Bidirektionaler asynchroner KS, Verbindungen mit hoher Bandbreite existieren zwischen beiden Plattformen
- Tier 5: Synchroner Update im jeweiligen Ausweich-RZ, kein Bandtransport mehr nötig
- Tier 6: Automatische Umschaltung im K-Fall

Wir verweisen im BS2000-Umfeld hier explizit auf drei exemplarische Typen von Verfahren zur Verfügbarkeit der Online-Daten im Standby-RZ:

In synchronen KS-Konfigurationen

1. Spiegelung der Daten mit synchronem SRDF (Online Spiegelung)
   Die Online-Daten der Produktionsanwendungen liegen auf einem oder mehreren Symmetrix-Systemen und werden mit synchronem SRDF über Verbindungsstrecken (Remote Links) zu (spiegelbildlichen) Symmetrix-Systemen im Standby-RZ gespiegelt.
   Die aktuellen Daten befinden sich stets (lokal geschützt) auf RAID1 Source Units im Produktions-RZ und darüber hinaus zusätzlich (sinnvollerweise auch dort lokal geschützt) auf RAID1 Target Units im Standby-RZ.
   Die fundamentale Aussage bei dieser Konfiguration ist die Tatsache, dass im K-Fall keine Daten verloren gehen.
   ➔ entspricht Tier 5-6 in IBM-Terminologie

In asynchronen KS-Konfigurationen

2. Spiegelung der Daten mit asynchronem SRDF (Online Spiegelung)
   Die Online-Daten der Produktionsanwendungen liegen auf einem oder mehreren Symmetrix-Systemen und werden mit asynchronem SRDF über Verbindungsstrecken (Remote Links) zu (spiegelbildlichen) Symmetrix-Systemen im Standby-RZ gespiegelt.
   Die aktuellen Daten befinden sich stets (lokal geschützt) auf RAID1 Source Units im Produktions-RZ. Den Target Units im Standby-RZ sind BCV Units zugeordnet, auf denen Konsistenzstände eingefroren werden können.
   Weil der Datenabgleich bei asynchronem SRDF-Betrieb nicht kontinuierlich (die Reihenfolge der Anwendungs-Schreib-IOs wird beim asynchronen SRDF nicht eingehalten) stattfindet, müssen für konsistente Daten im Produktions-RZ periodisch die Anwendung gestoppt und die SRDF-Volumes synchronisiert werden. Nach der Synchronisation und der Abtrennung der zugeordneten BCV-Spiegelvolumes von den Target-Volumes im Ausweich-RZ kann die Anwendung schon wieder gestartet werden.
   ➔ entspricht Tier 4 in IBM-Terminologie

3. Regelmäßiger Transport von Backup-Datenträgern (Offline Spiegelung)
   Bei dieser Variante werden die Daten nicht elektronisch, sondern physikalisch in Gestalt von Magnetband-Kassetten mit Daten-Sicherungen vom Produktions-RZ ins Standby-RZ gebracht. Es liegen zwangsläufig Ähnlichkeiten zu den Verfahren vor, wo die Daten-Sicherungen elektronisch von einem RZ zum andern gebracht werden.
   Der Austausch von physikalischen Bändern im allgemeinen RZ-Betrieb ist im Lauf der letzten zwanzig Jahre bewusst reduziert worden wegen seiner Nachteile bei Bandgeräteeinsatz und Bandverschnitt, Operatoreinsatz und schließlich wegen der damit erreichten mangelhaften Datenaktualität am Zielort. Mit der Verlagerung zu Datentransfer über Datenverbindungen einher ging die Entwicklung des Bandbetriebs zu Kassettenlaufwerken, zu Montierhilfen, schließlich zu virtualisierten Geräten und Volumes.
   Zusätzlich zum Sicherungsbetrieb müssen im Operatorbetrieb tägliche Duplizierläufe im Produktions-RZ und Restore-Läufe im Ausweich-RZ über direkt angeschlossene Magnetband-Kassetten-Systeme gefahren werden und mitsamt der zusätzlichen Transportlogistik dazwischen wird das Ergebnis bei der Datenaktualität am Zielort recht unbefriedigend.
   ➔ entspricht Tier 2-3 in IBM-Terminologie

Eine andere Variante, die dem Tier 1 entspricht, ist die Nutzung eines Dienstleistungsunternehmen (z.B. Restart), das erst im K-Fall ein passendes Equipment innerhalb weniger Tage liefert. Als Vorsorge werden lediglich Sicherungsbänder ausgelagert. Solche Anbieter können auch ein Symmetrix-System mit Target Units zur Datenspiegelung sowie VM-Gastsysteme und Extra-CPUs als Cold-Standby bereitstellen. Die CPU-Leistung wird dann erst im K-Fall aktiviert. Interessant ist diese Variante ggf. auch als Kooperationsmodell zwischen RZs unterschiedlicher Unternehmen.

3.4.1 Bewertungsaspekte für die Konfigurations-Varianten

A. Aktualität der Daten im Ausweich-RZ
B. Zeitdauer von der Ausrufung des K-Falles bis zum Wiederanlauf der Anwendungen
C. Kosten des additiv benötigten Personals und der additiv benötigten Betriebsmittel
D. Beeinflussung der Performance der Produktiv-Abläufe
E. Schutz der Daten gegen unbefugtes Lesen

F. Robustheit des Verfahrens

Zur Übersicht eine Beurteilungs-Matrix der Konfigurations-Varianten nach den oben genannten sieben Aspekten. Dabei bedeuten die Einträge einen vergleichenden Bewertungsrang: 1 = am besten, 3 = am wenigsten gut. Die einzelnen Aspekte haben unterschiedlichen Stellenwert.

	Synchr. SRDF	Asynchr. SRDF	Bandtransport
A. Datenaktualität	1	2	3
B. Ausfallzeit	2	1	1
C. Kosten	1	2	3
D. Performance	2	1	3
E. Datenschutz	1	1	3
F. Robustheit	1	2	3

Tab. 3-1 Beurteilungsmatrix für die Konfigurations-Varianten

Es liegt auf der Hand, dass die Variante 1 funktional hervorragend (= Hotstandby), und - wegen der durchgehend gleichzeitigen Aktualität der Daten in beiden Standorten, die bei den zwei anderen Varianten mitnichten gegeben ist - deutlich am besten zu bewerten ist.

Erläuterung zu den Bewertungen:

1. Datenaktualität
   Beim synchronen SRDF sind die Daten im Ausweich-RZ jederzeit auf dem gleichen Stand wie im Produktions-RZ. Für den K-Fall gibt es zwei Abstufungen: mit oder ohne Domino-Modus (2.8.2). Mit Domino-Modus sind die Daten in jedem denkbaren K-Fall 100%ig aktuell – dafür wird eine Abschwächung der Hochverfügbarkeit in Kauf genommen. Ohne Domino-Modus sind Rolling Disaster K-Fälle denkbar, bei denen die Daten im Ausweich-RZ zwar immer konsistent, aber u.U. nicht ganz aktuell sind (z.B. wenn die SRDF-Links zunächst ausfallen, und der K-Fall erst zu einem späteren Zeitpunkt Host und/oder das Symmetrix System lahm legt), siehe hierzu auch 2.8.2.
   Beim asynchronen SRDF hängt die Datenaktualität von der Häufigkeit des „Einfrierens“ eines Konsistenzstandes ab (siehe 4.4).
   Beim Bandtransport hängt die Datenaktualität von der Häufigkeit des Bandtransports ab – hier wird man wohl einen bestenfalls täglichen Rhythmus wählen.

2. Ausfallzeit
   Die Ausfallzeit ist bei den asynchronen Varianten etwas niedriger als bei der synchronen Variante, da auf die Target-Freischaltung verzichtet werden kann, wenn mit den BCVs im Standby-RZ weitergearbeitet wird. Näheres hierzu in Kap. 4.4.2 . Dieses bewerten wir jedoch bei weitem nicht so hoch, wie den Aspekt Datenaktualität.

3. Kosten
   Bei den ersten zwei Varianten Synchrone SRDF-Spiegelung und Asynchrone SRDF-Spiegelung fällt auf, dass wahrscheinlich die funktionell schwächere zweite Variante die etwas teurere ist, weil die notwendigen zusätzlichen BCVs zur Abspaltung (jeweils pro logischem Plattenpaar) wohl teurer sind als eine vielleicht etwas bessere Netzauslegung für Spitzenlast bei der ersten Variante. Ein Mix der beiden SRDF-Varianten – für einige Anwendungen und ihre Platten synchrone SRDF-Spiegelung und für andere asynchrone SRDF-Spiegelung – scheint nicht besonders sinnvoll, weil die Netzauslegung erschwert wird für kontinuierlichen Synchronbetrieb neben periodischem asynchronem Plattenabgleich.
   Nach den Kosten fällt die Variante Bandtransport aus dem Rahmen. Sie beansprucht zusätzliches Potenzial bei Bandtransport, Bandoperating, Bandmaterial und Bandgeräten. Auch nach absoluten laufenden Kosten fällt diese Variante deutlich aus dem Vergleich heraus, da sie die weitaus höchsten Personalkosten bedingt.

4. Performance
   Die Performance im laufenden Betrieb wird vom asynchronen SRDF am wenigsten belastet. Beim synchronen SRDF verlängern sich die IO-Zeiten für die Schreib-IOs (siehe 3.4.3). Bei der Variante Band-transport verändert sich die IO-Zeiten nicht, allerdings kommen aufwändige tägliche Restore-Zeiten am Standby-RZ additiv hinzu.

5. Datenschutz
   Mit Ausnahme der Variante "Bandtransport" ist bei allen genannten Varianten ein sehr hohes Maß an Datensicherheit bereits durch die Nutzung von Glasfaserverbindungen gegeben, die nur mit hohem technischen Aufwand "abhörbar" sind. Bei Nutzung der WDM Technologie (s. Kap. 3.4.2) müsste zusätzlich das Zeitmultiplexing (TDM) entschlüsselt und eine spezielle Wellenlänge (λ) herausgefiltert werden. Der potentielle Lauschangreifer bräuchte also ein eigenes WDM-Equipment und Kenntnis über das Equipment des Unternehmens. Auf Ebene des ESCON- oder FC-Protokolls entsteht außerdem eine zusätzliche Zerstückelung der Daten in maximal 2kB große Blöcke, die sich auf unterschiedliche Platten verteilen können. Bei der asynchronen Verbindung wird das Abhören durch die ungeordnete Reihenfolge der Daten auf dem SRDF-Link zusätzlich erschwert.
   Bei der Nutzung der WAN-Technologie (s. Kap. 0) ist das „Abhören“ der Verbindung prinzipiell einfacher – hier muss ggf. auf Channel Extender mit encryption-Funktionalität zurückgegriffen werden.

6. Robustheit
   Durch ihre erforderlichen additiven K-Fall-Vorsorge-Aktivitäten wird vor allem bestimmt, wie robust oder fehleranfällig die Verfahren sind. Besonders robust ist die erste Variante, weil hier nach der einmaligen Aufnahme der synchronen SRDF-Spiegelung keine regelmäßige Bedienung mehr notwendig ist. Die zweite Variante ist ebenso robust, weil der

Datenabgleich automatisiert werden kann. Dieser Automatismus muss jedoch gepflegt werden, wobei natürlich Fehler möglich sind. Die dritte Variante ist nicht zu automatisieren, und somit ist täglich die Möglichkeit von Fehlern gegeben.

Durch diese Ausführungen wollen wir deutlich machen, dass die asynchronen Verfahren in nahezu jeder Hinsicht den synchronen Verfahren unterlegen sind. Im nächsten Abschnitt sollen KS-Konfigurationen diskutiert werden. Wir beschränken uns deshalb fast ausschließlich auf synchrone KS-Konfigurationen und behandeln von den asynchronen KS-Konfigurationen lediglich die Spiegelung der Daten mit asynchronem SRDF ausführlicher.

3.4.2 Remote Link Varianten bei großen Entfernungen

Um eine SRDF-Verbindung über eine weite Entfernung (Erläuterungen zu den Größenordnungen unter Kap. 3.5, „entfernungsabhängige Einschränkungen“) führen zu können, benötigt man

- entweder eine WAN-Netzverbindung zwischen den beiden Symmetrix-Standorten und an beiden Standorten je einen **Channel Extender**, der das SRDF-Protokoll auf ein Netz-Protokoll (T3, ATM, etc) umsetzt und somit die Verbindung zwischen Symmetrix und dem Netzanschluss herstellt, alternativ eine TCP/IP Netzverbindung und statt einem Channel Extender ein „Storage to IP Gateway“
- oder eine Lichtwellenleiter-Verbindung (Point-to-Point) über die gesamte Strecke zwischen den beiden Symmetrix-Systemen und die **WDM-Technik**, grundlegende Einführung dazu beispielsweise in [7].

WAN-Netzverbindung
Für eine WAN-Netzverbindung der über SRDF zu koppelnden Symmetrix Subsysteme wird eine Netzverbindung zwischen den beiden Standorten und pro Standort ein Protokollumsetzer von ESCON oder FC auf das Netzprotokoll (T3, ATM, etc), ein so genannter Channel Extender benötigt. Hersteller solcher Channel Extender sind z.B. die Firmen Inrange Technologies mit ihrem Storage Network System „9801 SNS“ und Computer Network Technology (CNT) mit den Produkten aus der „UltraNet Product Family“ oder „Channelink Product Family“. Die zu übertragenden Daten werden in den Channel Extenders komprimiert und mit dem entsprechenden Netzprotokoll verschickt. Das Netz muss entsprechend dem Datenstrom breitbandig ausgelegt sein und sollte für die SRDF-Datenübertragung exklusiv zur Verfügung stehen. Die Dauer einer einzelnen Schreib-IO wird verlängert um die Umsetz-Zeiten in den Channel Extendern und um die mit wachsender Entfernung steigenden Netzlaufzeiten sowie insbesondere durch die Verwendung mehrerer Router oder ATM-Switches auf dem Netzweg. Weiter können Daten-Kompressions- und Daten-Verschlüsselungs-Verfahren im Channel-Extender noch einen additiven Zeit-Aufschlag verursachen.

Die anfallenden Kosten bestehen in der Anschaffung der Channel Extender und den Kosten für die permanente, exklusive und breitbandige Netzverbindung (mindestens T3 34 Mbit/s, besser ATM 155 Mbit/s oder höher).

Bei der Festlegung der Ausgestaltung der WAN-Netzverbindung stehen verschiedene Technologien und HW-Hersteller sowie verschiedene Netzwerkbetreiber, Telekommunikationsanbieter und Provider mit unterschiedllichen Geschäfts- und Servicemodellen zur Auswahl. Fujitsu Technology Solutions wird bei dieser Auswahl gerne beratend und unterstützend mitwirken.

WDM-Technologie
Durch den Einsatz von Multiplexverfahren können vorhandene Übertragungsleitungen mehrfach und daher wirtschaftlicher genutzt werden. Bei Telefonnetzen haben sich seit langem Zeitmultiplexverfahren (Time Division Multiplexing, TDM) durchgesetzt. Hierbei geht es um die Aufteilung der einzelnen Verbindungen in Zeitschlitze, jedem Übertragungskanal wird ein fester Zeittakt zugeteilt.

Das heute optimale technische Verfahren für eine SRDF-Verbindung über weite Entfernungen ist jedoch die Nutzung der Wavelength Division Multiplex (WDM) Technologie oder einer Kombination dieser beiden Multiplex-Techniken WDM und TDM:

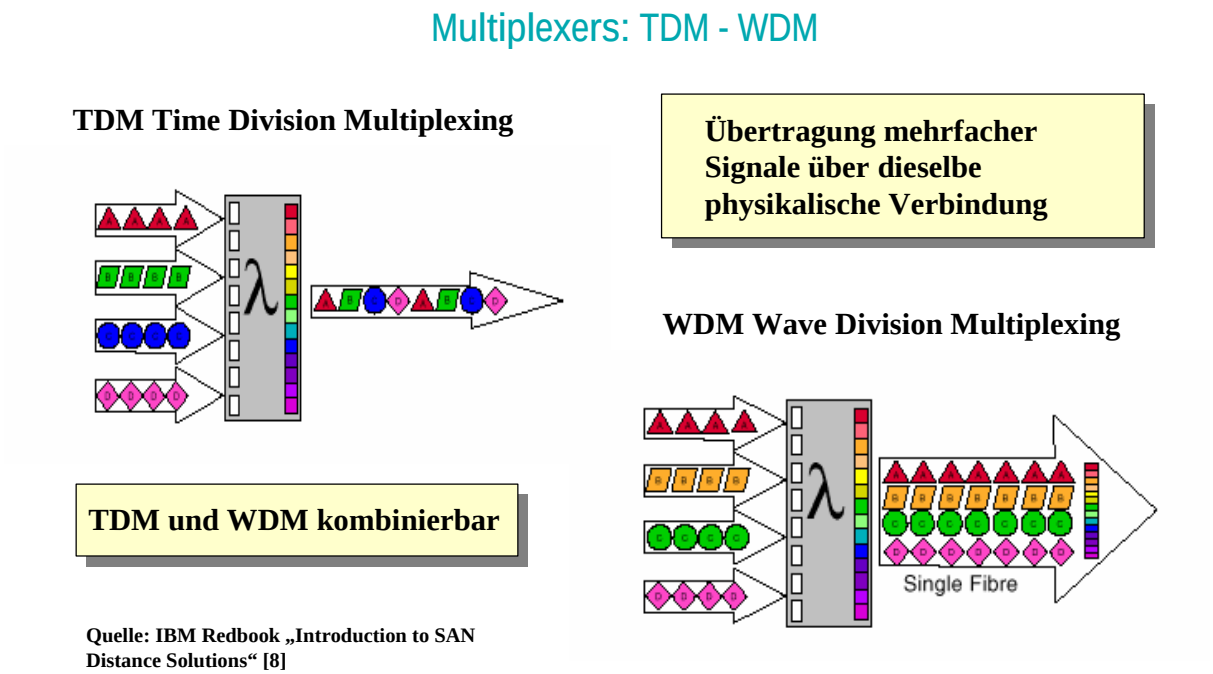

Multiplexers: TDM - WDM
TDM Time Division Multiplexing
Übertragung mehrfacher Signale über dieselbe physikalische Verbindung
WDM Wave Division Multiplexing
TDM und WDM kombinierbar
Single Fibre
Quelle: IBM Redbook „Introduction to SAN Distance Solutions“ [8]

Abb. 3-2 Multiplexers TDM und WDM

Wavelength Division Multiplex ist ein optisches Verfahren, welches mit rein optischen Komponenten funktioniert. Verwendet wird Glasfaserkabel, welches Laserlicht überträgt. Der Clou an der Sache ist, dass das Laserlicht Licht unterschiedlicher Wellenlänge beinhaltet. Auf der Senderseite werden die optischen Signale (FC, ESCON oder sonstige) in elektrische umgewandelt, dann auf jeweils einen Laser bestimmter Wellenlänge moduliert und mit den anderen Lasern zusammengeführt (multiplexiert) und anschließend in einer einzigen Glasfaser übertragen, wie in Abb. 3-2 gezeigt. Auf der Empfängerseite geschieht die umgekehrte Operation, indem die einzelnen Signale wieder voneinander getrennt (demultiplexiert) und dann ihren entsprechenden Empfängerdioden zugeführt werden. Auf diese Art können über nur eine Glasfaser 64 (demnächst sogar 128) bidirektionale Kanäle und über ein Glasfaserkabel eine sehr große Anzahl Kanäle (z.Z. bis zu 4000), die sich aufgrund ihrer unterschiedlichen Wellenlänge nicht beeinflussen, realisiert werden.

Die einzelnen Kanäle, die sich durch verschiedene Wellenlänge unterscheiden, werden auch λ‘s („Lambdas“) genannt. Dieses optische Verfahren erlaubt es also, über eine sehr begrenzte Anzahl von Kabeln eine sehr hohe Menge an Daten zu übertragen. WDM besitzt das Potential, die existierende Glasfaser-Netzinfrastruktur optimal auszunutzen. Ein weiterer Vorteil besteht darin, dass die Laser-Signale auf rein optischer Basis verstärkt werden können, falls größere Entfernungen zu überbrücken sind. Hierfür gibt es Erbiumdotierte Faserverstärker (kurz: EDFA = "Erbium Doped Fiber Amplifier“), die nach einer Entfernung von 60 bis 150 km (abhängig vom genutzten WDM-Equipment und der Glasfaser) einzusetzen sind. Hierdurch spart man sich die zeitliche Verzögerung durch Umwandlung des optischen Signals in ein elektrisches und wieder zurück und somit evtl. Performance-Einbußen. Bei geringer Dämpfung der Glasfaser (db/km) können auch mehrere 100 km ohne elektrische Regenerierung des Lichtsignals überbrückt werden.

Neben der Nutzung eines λ‘s als ESCON-Kanal durch Verwendung einer ESCON-Kanalkarte im WDM-Gerät können auch über entsprechende weitere Karten im WDM-Gerät FC, ATM, Fast-Ethernet, Voice etc. über andere λ‘s genutzt werden. So könnte z.B. ein gesamtes Corporate Network über WDM-Technologie abgewickelt werden, (s.a. Kap. 3.3) da auch optisches Routing möglich ist (mit dem optischen Add-Drop-Multiplexer gibt es die Möglichkeit, optische Ringtopologien zu bauen).

Die anfallenden Kosten bestehen in Anmietung, Kauf oder Leasing einzelner WDM-kompatibler Glasfasern bzw. einzelner λ‘s sowie den erforderlichen WDM-Geräten. Manche Provider bieten als plugin Lösungen ESCON- oder FC-Kanäle und „managed service“ an. Die Kosten verglichen mit der WAN-Netzverbindung gestalten sich bei WDM eher günstiger bei zudem deutlich besserer Funktion und Leistung.

Aus all diesen Gründen (leistungsfähiger, zukunftsträchtiger, kostengünstiger) geben wir der WDM-Technologie auch gegenüber der WAN-Netzverbindung den eindeutigen Vorrang.

Kurz wesentliche technische Eigenschaften auf einen Blick:

WDM = Wave Division Multiplexing, DWDM = Dense Wave Division Multiplexing
- Monomode Lasertechnik
- 9 Micrometer Durchmesser

- bis zu 64 (128) Anwendungen pro Verbindung (DWDM)
- 64 (128) x 2,5 (10) GigaBits Bandbreite einer Glasfaser
- bis zu 150 km ohne Verstärkung
- Verstärkung rein optisch

WDM-Technik ist einsetzbar für
- EMC² SRDF (ESCON mit FarPoint)
- Fibre Channels
- Shared DASD
- Remote Peripherie
- CTC-Connections
- Fast-Ethernet, Gigabit-Ethernet, ATM

Die Performance einer synchronen SRDF-ESCON-Verbindung wird durch die EMC-Firmware „Far Point“ verbessert, die eine parallele Abarbeitung mehrerer IOs für unterschiedliche Volumes über die Remote Links erlaubt. Far Point ist allerdings nur einsetzbar für eine unidirektionale SRDF-Verbindung. Bei einer bidirektionalen SRDF-Verbindung müssen, um von der Performance-Verbesserung durch Far Point profitieren zu können, die Remote Link Verbindungen zu zwei unidirektionalen sogenannten RA-Gruppen (remote adapter groups) bei der Installation gruppiert werden.
Um also eine besonders performante Remote Link Verbindung über ESCON zu realisieren, sind mehrfache Remote Links und damit auch mehrfache ESCON-Boards erforderlich.

Die Performance einer synchronen SRDF-FC-Verbindung wird durch die Entfernung, die verwendeten Switches und deren maximale „Buffer-to-Buffer Credit Rate“ bestimmt. Sie übersteigt die ESCON-Performance bei „kleineren“ Entfernungen, wobei der „breakeven point“ bei den Entfernungen für vorgegebene Konfigurationen berechnet werden kann. Beispielsweise bei Verwendung von Switches mit 1 Gbit/s (damit beträgt die Länge eines Datenframes in der Glasfaserstrecke 4 km) und mit einer eingestellten Buffer-to-Buffer Credit Rate von 60 (s. hierzu auch Kap. 3.5) hat der FC Link bei Entfernungen bis 4 km x 60 = 240 km (bzw. einfache Strecke 120 km) Performance-Vorteile gegenüber dem ESCON Link.

Das benötigte WDM-Equipment wird von vielen Herstellern bereitgestellt resp. vertrieben. Eine beispielhafte, aktuelle Auswahl umfasst (baugleiche Produkte in einer Zeile zusammengefasst):
- Siemens (Waveline), CNT/Inrange (Spectrum), ADVA (FSP), Alcatel (Optinex)
- Sorrento Networks (GigaMux), Inrange (GigaMux)
- Nortel Network (OPTera Metro)
- Cisco (Metro, ONS)
- CNT (UltraNet Wave Multiplexer)
- ONI (Online)

Bei der Festlegung der Ausgestaltung der WDM-Verbindung stehen verschiedene Technologien und HW-Hersteller sowie verschiedene Glasfaser-Netzwerkbetreiber, Telekommunikationsanbieter und Provider mit unterschiedllichen Geschäfts- und Servicemodellen zur Auswahl. Fujitsu Technology Solutions wird bei dieser Auswahl gerne beratend und unterstützend mitwirken.

3.4.3 IO-Verteilung

Es ist klar, dass bei Nutzung von synchronem SRDF über weite Entfernungen jede Schreib-IO eine gewisse Verzögerung erfährt. Dies sei anhand des folgenden Beispiels kurz erläutert. Die Schreib-IO ist bei synchronem SRDF beendet, sobald die Daten auch im Cache der entfernten Symmetrix eingelagert und ein Quittierungs-Signal zurückgegangen ist. Der Zeitverlauf einer Schreib-IO über den SRDF-Link und die entfernte Steuerung setzt sich dann wie folgt zusammen, wobei wir im Beispiel von einer ESCON-SRDF-Verbindung ausgehen, die über WDM-Multiplexer geführt wird, die Entfernung sei 250 km:

t_{local}	Zeit, die die lokale Symmetrix braucht, um den Auftrag ins Cache einzulagern, wir nehmen hier eine Zahl von 0.5 ms an.
t_{prot}	Zeit, die benötigt wird, um aus dem opt. ESCON ein WDM-Lasersignal zu machen.
t_{data}	Zeit, die ein Datenblock von 4 KB über den Link benötigt, diese ist im Beispiel vorgegeben durch die Übertragungsrate von ESCON (SRDF) und betrage 0.3 ms. (hypothetischer Wert, der Minimalwert ist bei 200Mbit/s 0,2ms)
t_{line}	Signallaufzeit, die durch 250 km Glasfaserkabel benötigt wird, bei der Ausbreitungsgeschwindigkeit 200 000 km/s von Licht in Glasfaser sind dies bei 250 km 1.25 ms
t_{remote}	Zeit, die die entfernte Symmetrix braucht, um den Auftrag ins Cache einzulagern und zu quittieren, wir nehmen hier eine Zahl von 0.5 ms an.

Da die Umsetzung viermal erfolgt (je einmal in jede Richtung an beiden Standorten) und der Weg durch die WDM-Glasfaser zweimal durchlaufen werden muss, (vom Datenblock und vom Quittierungs-Signal) kann man die Verzögerungszeit durch SRDF über WDM wie folgt berechnen:

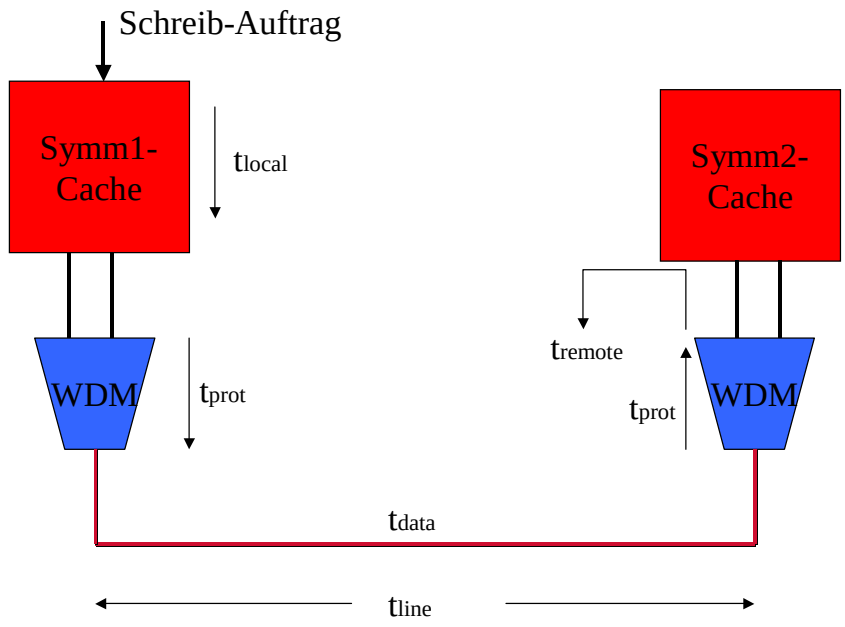


Abb. 3-3 Verlauf einer Schreib-IO bei synch. SRDF

- $t_{sum} = t_{local} + t_{delay} + t_{data} + t_{remote}$ (Gesamt-Zeit einer 4KB-Schreib-IO)
- $t_{delay} = 4 * t_{prot} + 2 * t_{line}$ (Verzögerung durch die Entfernung des SRDF-Links)

Selbst wenn man die Umsetzzeit der WDM-Multiplexer beliebig klein annimmt ($t_{prot} = 0$) , kommt man auf eine Gesamtverzögerung durch die weite Entfernung von

- $t_{delay} = 2 * 1.25$ ms ~ 2.5 ms

und damit zu Gesamtzeiten von

- $t_{sum} = 0.5$ ms + 0.3 ms + 0.5 ms = 1.3 ms bei Entfernung 0 m
  und
- $t_{sum} = 0.5$ ms + 2.5 ms + 0. 3 ms + 0.5 ms = 3.8 ms bei Entfernung 250 km

Wenn man von einer einzelnen Platte und synchronen Schreib-IOs ausgeht, so sind damit
- bei Entfernung 0 bis zu 806 Schreib-IOs/s möglich,
- bei Entfernung 250 km aber nur 1000/3.74 ~ 267 Schreib-IOs/s.

Vergleich: mit **einem** PKW kann man in begrenztem Zeitraum auch nur eine begrenzte Anzahl Personen über 250 km transportieren, selbst wenn die Autobahn 10-spurig ist und ein hohes Tempo erlaubt.

Werden viele Platten parallel genutzt, kann natürlich trotzdem die maximale Datenrate auf der SRDF-Verbindung erreicht werden. Deshalb ist vor Einsatz einer SRDF-Verbindung über weite Entfernung zu untersuchen, ob es sogenannte „Hotspot"-Anwendungen gibt, die temporär hohe Anforderungen an Durchsatz von Schreib-IOs auf eine Platte stellen.

Für Kunden, die eine KS-Konfiguration mit SRDF über längere Distanzen planen, bietet EMC ein „SRDF-Distance-Kit". Hierin enthalten sind zum einen die Analyse der bereits vorhandenen Platten auf ihre Auslastung und die Ermittlung der Hotspots sowie die Errechnung der benötigten Bandbreite über (D)WDM. Weiterhin gibt es ein Testequipment, das im wesentlichen aus WDM Mux/ Demux von Sorrento Networks oder Inrange Technologies und aus 6 Glasfaser-Kabelrollen mit Zusatzhardware besteht. Hiermit kann eine „long-distance"- SRDF-Verbindung auf Basis von ESCON oder FC simuliert werden. Mit den enthaltenen Glasfaserkabeln können Entfernungen bis zu 100 km (bei 4 Links) oder 200 km (bei 2 Links) in Schritten von 10 km eingerichtet werden. Somit kann verifiziert werden, ob die erforderliche Gesamt-IO-Rate auch über die geforderte Entfernung noch erreicht wird.

3.5 Entfernungsabhängige Einschränkungen

In vielen Dokumentationen über KS werden die Konfigurationen oder Cluster in drei Klassen mit unterschiedlichen räumlichen Ausdehnungen eingeteilt:
- Campus-/ Nah-Bereich: Ausdehnung bis max. einige km.
- Metropolitan-Bereich: Ausdehnung bis in einen Bereich von ca. 100 km
- Continental-Bereich: Ausdehnung bis zu mehreren 100 km.

Mit den für BS2000-Systeme derzeit relevanten ESCON- oder FC-Protokollen können Campuslösungen ohne zusätzliche Maßnahmen realisiert werden, solange zwischen den RZs Glasfasern beliebig verlegt werden können. (d.h. insbesondere bei einem echten Firmen-Campus, wo keine Leitungen über öffentlichen Grund verlegt werden müssen). Größere Ausdehnungen im Metropolitan -Bereich werden i.a. die Beauftragung eines Netzcarriers erfordern, der die benötigte Bandbreite (mit akzeptabler Laufzeit) auf einer oder mehreren Glasfasern („Dark Fibre") zur Verfügung stellt. Für diesen Bereich bietet sich die WDM-Technologie an. Gleiches gilt für RZs, die nur wenige km voneinander entfernt sind, die aber nicht über firmeneigenes Gelände miteinander verbunden werden können. Für eine Verbindung im Continental-Bereich wird man ggf. eine wie unter Kap. 3.4.2 besprochene WAN-Verbindung heranziehen müssen, wobei die Latenzzeiten einzelner IOs sowie die entstehenden Kosten dann kritisch zu prüfen sind.

Auf der physikalischen Ebene ist zunächst der Lichtwellenleiter-(LWL) Typ zu wählen. Angaben hierzu sind in der einschlägigen Fachliteratur zu finden. Mit den heutigen FC-Verbindungen kommt man hier auf maximal 10 km zwischen zwei Ports. ESCON Verbindungen können bis zu 3 km betragen, Verbindungen zwischen zwei ESCON-Direktoren (SCDs) 20 km. Durch die Verbindungen zwischen SCD und Host und die zwischen SCD und Gerätesteuerung kann eine ESCON-Strecke so auf 26 km ausgedehnt werden.
Diese Entfernungsgrenzen können bei Einsatz von WDM-Strecken überwunden werden, hier sind derzeit bis zu 100 km ohne Verstärkung des Lichtsignals möglich, bei Einsatz optischer Verstärker etwa 200 km. Bei noch größeren Distanzen ist dann abzuwägen, ob eine ATM oder TCP/ IP – Verbindung in Frage kommt, diese Fragen wird man dann jedoch mit einem Netzcarrier oder einem Anbieter von Komplettlösungen klären müssen.

Performance-bedingte Grenzen können sich auf der Protokollebene von ESCON oder FC ergeben. Diese sind zwar über beliebige Entfernungen funktionsfähig, jedoch wird die maximale Übertragungsrate zwischen zwei Komponenten bei gewissen Distanzen nicht mehr erreicht. Ein Einbruch der Performance, auch „droop" genannt, setzt ein, sobald die Laufzeit eines Lichtsignals im LWL, aufgrund der Entfernung zwischen Sender und Empfänger, die Größenordnung der Übermittlungsdauer eines Datenframes erreicht. Für den Sender ergeben sich dann relativ lange Wartezeiten auf das Quittungssignal, und dadurch sinkt die effektive Datenrate auf der Verbindung. Bei ESCON beispielsweise beträgt die Übertragungsrate 200 Mbit/s und damit wird ein 2KB, (also 20Kbit) großer Frame in 100µs übertragen. Die Ausbreitungsgeschwindigkeit des Lichts beträgt in Glasfasern 200 000 km/s, d.h. in diesen 100µs legt es 10 km zurück. Somit ist bei ESCON der Sender bei einer Entfernung von

20 km Hin- und Rückweg, also 10 km einfacher Entfernung, weitere 100µs im Wartezustand auf das Quittungssignal. Er sendet also nur mit der halben Leistung (diese Darstellung ist allerdings vereinfacht, da ein 2KB-Block am ESCON-Kanal je nach Architektur der Plattensteuerung in kleineren Portionen, jeweils mit eigenem Quittungssignal, übertragen wird. Diese Wartezeiten können jedoch durch die Möglichkeit, mehrere Frames zu senden, ohne auf ein Bestätigungssignal zu warten, reduziert werden. Bei FC gibt es dafür die so genannten „Buffer-to-Buffer Credits“. Der droop beginnt nach 9 km, und bei 23 km wird nur noch die Hälfte des Durchsatzes erzielt. Bei FC liegt diese Grenze wegen der hohen Datenrate von 1Gbit/s resp. 2Gbit/s bereits bei 4 km resp. 2 km, diese kann allerdings durch die Buffer Credits der eingesetzten Fibre Channel Switches oder FC/IP Gateways derzeit auf jeweils ca. 120 km ausgedehnt werden. Je nach Anforderungen der Anwendungen ist es abzuwägen, wie groß die zusätzlichen, entfernungsabhängigen Laufzeiten sein dürfen.

Bei Einsatz von ESCON mit Farpoint für die SRDF-Links setzt der o.g. Abfall deutlich später ein, bei Messungen mit mehreren ms simulierter Verzögerungszeit (1ms Verzögerung entsprechen 200 km) haben wir nur einen leichten linearen Abfall der Transferraten feststellen können. Die Verbindungen von Host zu Host, die in unseren Betrachtungen im Wesentlichen von BCAM und MSCF genutzt werden, sind am wenigsten zeitkritisch, da hier keine hohe Last erzeugt wird.

Wenn nicht weiter erläutert, beschränken wir in folgenden Beschreibungen den „Campusbereich“ auf 10 km, den „Metropolitan-Bereich“ auf ca. 100 km. Alles darüber hinaus gehende wird von uns als „Continental-Bereich“ bezeichnet.

Zusammenfassung der Maximal-Entfernungen:

Perf. Host-Peripherie-Verbindung bei ESCON-Anschluss:	bis zu 9 km
Perf. Host-Peripherie-Verbindung bei FC-Anschluss:	bis zu 120 km
Perf. SRDF-Verbindung über ESCON und FarPoint:	bis zu mehreren 100 km
Perf. SRDF-Verbindung über FC und FC-Switches:	bis zu 120 km
Perf. SRDF-Verbindung über FC und FCIP-Gateways	bis zu 200 km

Dabei ist die SRDF-Verbindung mit ESCON-Directors und FarPoint erst ab 120 km (dem heutigen “Einbruchspunkt” von FC) der SRDF-Verbindung mit FC-Directors überlegen.

Alle diese Aussagen sind relativ zum heutigen Technik-Stand zu interpretieren, also keine Absolut-Werte.

3.6 Administrative Vorbereitungen für die Katastrophenvorsorge

Außer den in den vorigen Kapiteln genannten Maßnahmen sind auch diverse administrative Vorbereitungen zu treffen, um einen möglichst reibungslosen Ablauf im K-Fall sicherzustellen; es folgt eine exemplarische Liste ohne Anspruch auf Vollständigkeit:

- Eintragen der im K-Fall umzuschaltenden Userkennungen am Ausweich-Host
- Eintragen der Hostnamen / IP-Adressen der Anwender der Umschalteinheit (auch bei Einsatz von VM2000).
- Anlegen der MRSCAT-Einträge aller umzuschaltender Datenpubsets am Ausweich-Host.
- Falls mit OMNIS gearbeitet wird, müssen die OMNIS-Kennungen auch am Ausweich-System eingerichtet werden.
- Erstellen von erweiterten Netz-Konfigurationsdateien (SOF, RDF) für die verschiedenen Standorte und ggf. Einrichten der virtuellen Hosts.

Wenn VM2000 zur Trennung der Anwendungen genutzt wird (s.a. Kap. 4.7), ist folgendes zu beachten:
- Es sind am Standby-RZ die Spiegel-Gastsysteme einzurichten, denen die ermittelten Werte (des Produktiv-RZ) für CPU- und Hauptspeicherbedarf zugewiesen werden.
- Die Netzadressen der Gastsysteme für die Umschalteinheiten müssen im Monitorsystem eingetragen werden und es müssen virtuelle Konsolen oder SKP-Konsolen für die zusätzlichen Gastsysteme an den Monitorsystemen eingerichtet werden.
- Die im Standby-RZ eingerichteten Platten für Daten- und Home Pubsets werden der VM zugewiesen, ebenso die Netzzugänge (HNC) und die virtuellen Konsolen.

Ist HIPLEX AF im Einsatz, dann müssen die dafür nötigen Vorbereitungen getroffen werden, wie
- Einrichtung eines MSCF-Verbundes
- Zusätzlich Einrichtung eines Shared-Pubset-Verbunds bei X-Link-Konfigurationen
- Definition der Switch-Units für die Anwendungen
- Erstellen bzw. Anpassen von Failover-/ Failback-Phasen.

Diese Vorbereitungen sind in den Handbüchern [3] und [4] beschrieben und sollen hier nicht weiter erläutert werden.

3.6.1 Behandlung von System und Paging Volumes

System Volumes
Bei einer KS-Konfiguration besteht die Möglichkeit, die System-Volumes genau wie die Anwendungs-Volumes per SRDF zu spiegeln und die Systeme erst im K-Fall im Ausweich-RZ zu starten oder Systeme am Ausweichstandort in einem VM2000-Gastsystemrahmen bereitzuhalten, die eine Anwendung im K-Fall übernehmen. Oder aber alle Anwendungen laufen unter einem Native-System, auch im K-Fall.

Bei der Variante „gespiegelte Home Pubsets“ entstehen folgende Nachteile:
- im K-Fall geht bis zum Start der Anwendung zusätzlich auch die Zeit zum Hochfahren des Systems des Systems ein,
- im Normalbetrieb kann ein Standby-System erst nach SRDF-Trennung genutzt werden,
- auch an die Session gebundene Dateien (etwa Konsol-Logging) werden über SRDF kopiert, ohne dass sie im Home Pubset des Standby-Systems gebraucht werden,

- alle für das Standby-System spezifischen Dateien (etwa Netzkonfigurationsdateien) müssen schon am Work-System eingebracht werden,
- eine spezielle Auswahl-Logik (über unterschiedliche Parameterdateien) ist notwendig, damit von den speziellen Dateien automatisch jeweils die richtige ausgewählt wird.

Dieses Modell hat die folgenden Vorteile:
- Updates von allgemeinen Dateien (etwa Replader) werden automatisch auf beiden Systemen wirksam,
- Änderungen von Systemeinstellungen, die in Systemdateien geführt werden (etwa neue Kennung mit Default-CatID), sind automatisch auch für das Ausweich-System wirksam.

Das Alternativmodell dazu wäre ein Paar von gleichartig aufgebauten, aber getrennten Home Pubsets (ohne SRDF-Spiegelung). Die Liste der Vor- und Nachteile ist praktisch umgekehrt zum Spiegelmodell; beide Modelle können nicht sinnvoll kombiniert werden. Bei unabhängigen Home Pubsets müsste ein Update von allgemeinen Dateien z.B. per Filetransfer auf beiden Home Pubsets geführt werden (die Erprobung etwa von neuen Repladern zuerst am Standby-System wäre auch ein Vorteil) und einige Systemkommandos müssten zusätzlich für die Standby-VM gegeben werden. In jedem Fall muss bei unabhängigen Home Pubsets vorausgesetzt werden, dass das Standby-System im Normalfall schon läuft (evtl. mit reduziertem Hauptspeicher) und vom gleichen Systemverwalter betreut wird wie das Original-System.

Paging Volumes
Das Spiegeln von Pagingdaten bringt keinen Vorteil, und somit sollten Pagingdateien nach Möglichkeit auf Volumes angelegt werden, die nicht gespiegelt sind, also auch keine geschäftskritischen Daten enthalten, oder auf dem Homepubset, falls dieses nicht gespiegelt wird. Außer der Ersparnis an Plattenspeicher und der Vermeidung von unnötiger Last auf dem SRDF-Link bringt dies den Vorteil, dass das Hochfahren der Systeme nicht zusätzlich abhängig ist vom SRDF-Zustand der jeweiligen Source- bzw. Target Unit.

3.7 Datensicherungs-Konzept

Neben den Online-Daten gehören die Archivdaten zum Datenbestand einer Anwendung, d.h. diejenigen Daten über bereits abgeschlossene Geschäftsvorgänge, welche aus rechtlichen Gründen über längere Zeit aufgehoben werden müssen, z.B. für Reklamationen oder zur Rekonstruktion vergangener Geschäftsabläufe bei Verdacht auf Unregelmäßigkeiten. Diese Archivdaten müssen ebenfalls im Standby-RZ verfügbar sein, um in einem K-Fall den Geschäftsbetrieb ordnungsgemäß fortsetzen zu können.

Eine grundsätzliche Anforderung an jedes Backup-Konzept im Rahmen von Katastrophenschutz-Konzepten jedweder Ausprägung ist deshalb die Abschottung des Aufbewahrungsortes der Sicherungs-Datenträger. Dieser Aufbewahrungsort kann ein Brandschutz-Kellerraum sein, oder einfach ein vom Aufstellungsort der Server und Storage Systeme räumlich möglichst weit entfernter und gegen vielfältige Bedrohungen wie Feuer, Wassereinbruch, Explosionen, Diebstahl, Sabotage u.s.w. speziell gesicherter Ort. Allgemein gesprochen ist ein Konzept eines Remote Backup in jedem Fall empfehlenswert. Schließlich ist die Möglichkeit des Zugriffs auf die Sicherungsdatenträger auch die ultima ratio bei „logischen Fehlern“, nämlich dann wenn durch nie auszuschließende menschliche Fehler die aktuellen Produktionsdaten (und bei Spiegelungs-Konzepten damit auch die gespiegelten Daten in einem Ausweich-RZ) unbrauchbar werden.

Die Bandsicherungen werden bei den von uns vorgestellten Konfigurationen nicht für den Wiederanlauf der Anwendungen benötigt (außer bei einem speziellen Rolling Disaster, s. Kap. 2.8.2) – daher ist die Planung dessen, was am Standby-Standort an Sicherungsperipherie und gespiegelten Datenbeständen vorgehalten wird, abhängig von der Notwendigkeit, auch auf Langzeitsicherungen im Standby-RZ jederzeit zugreifen zu können.

Grundsätzlich sind mehrere Konzepte für die Datensicherung denkbar, um auch die Archive redundant zu halten:
1. Sicherungsbestände werden mit Bandtransfer und Bandduplikaten in ein zweites Archiv, was beim K-Fall unbetroffen wäre, verlagert. Dieses kann natürlich auch am Ausweichstandort liegen. Es müssen dafür die Kassettensysteme, die im Work-RZ genutzt werden, auch im Standby-RZ vorhanden sein. Allerdings entsteht personeller Aufwand durch den Bandtransport. Es müssen am Ausweichstandort die HSMS-Archive eingerichtet - und möglichst auch gepflegt – werden, um im K-Fall die Daten der Langzeitsicherung am Ausweichstandort einspielen zu können.
2. Es werden auch am Ausweichstandort Bandsicherungen durchgeführt. Zu den Zeiten der Sicherung am Work-RZ, an denen die Anwendung(en) nicht läuft, wird die SRDF-Spiegelung für die Datenpubsets per SHC-OSD-Kommando beendet, dann die Target Units in der Standby-Symmetrix freigeschaltet und die Pubsets – auf den Target Units- dort ebenfalls importiert. Dann wird hier die gleiche Sicherung wie am Work-RZ durchgeführt.

Falls es keine genügend langen Sicherungszeitfenster gibt, hat man die Möglichkeit über Additional-Mirror-Units (BCVs) zu sichern (s.a. Kap. 2.9 und [1]). Dies kann gleichzeitig im Work- und Standby-RZ erfolgen. Zu definierten Konsistenzzeitpunkten werden die BCVs von den Source und Target Units getrennt und auf Band gesichert. Hiermit entsteht allerdings zusätzlicher Plattenaufwand. Das Sicherungsverfahren bedeutet zusätzliche Routinearbeit, die jedoch automatisiert werden kann.

Datei-Migration
Wir empfehlen, auf Datei-Migration auf Bänder (S2) für geschäftskritische Daten zu verzichten. Wenn dies jedoch nötig ist, sollte dafür gesorgt werden, dass die Dateien zuvor auch in ein gesichertes Backuparchiv gelangen.

3.8 Stufenweise Einrichtung der KS-Voraussetzungen

Da die in diesem Kapitel genannten Voraussetzungen und Vorbereitungen bei hohen Rechner-, Plattenspeicher- und Netzkapazitäten auch hohe Kosten und Aufwände bedeuten können, kann eine stufenweise Einführung eines KS-Konzeptes angebracht sein. So kann beispielsweise in einem ersten Schritt ein Standby-RZ mit der erforderlichen Host- und Plattenkapazität ausgerüstet werden, um dort im K-Fall mit Hilfe von Sicherungskassetten schnell wieder aufsetzen zu können, um erst in einem späteren Schritt eine Datenspiegelung auf Plattenebene (SRDF) einzuführen, und um wiederum ggf. in einem weiteren Schritt das Datennetz, falls nötig, auszubauen.

Oder man beginnt mit einer lokalen Datenspiegelung, bei der die SRDF Target Units in einem Symmetrix Subsystem untergebracht sind, dass vom Original Symmetrix Subsystem mit den Source Units räumlich separiert ist. In einem späteren Schritt kann dann z.B. die Einrichtung einer WDM-Infrastruktur und wiederum später der Umzug der Symmetrix Spiegel Systeme erfolgen. Hier sind offensichtlich unterschiedliche Reihenfolgen denkbar und eine genaue Planung erforderlich.

4 KS-Konfigurationen

In diesem Abschnitt sollen verschiedene empfehlenswerte KS-Konfigurationen unter dem Aspekt der Verfügbarkeit und Aktualität der Online-Daten diskutiert werden. Wir beschränken uns fast ausschließlich auf synchrone KS-Konfigurationen und behandeln von den asynchronen KS-Konfigurationen lediglich die Spiegelung der Daten mit asynchronem SRDF ausführlicher. Nur die Konzepte mit synchroner Datenspiegelung (real time mirroring) bieten die Möglichkeit, im K-Fall mit aktuellen Daten den Betrieb in kurzer Zeit wiederaufnehmen zu können. Das Wiederaufsetzen auf Sicherungsbänder bedeutet dagegen in aller Regel den Verlust der Daten eines Tages und ein zeitintensives Restaurieren der Daten.

Alle Varianten beziehen sich auf eine Spiegelung der Daten mit SRDF. Die Unterschiede ergeben sich durch den SRDF-Modus (synchron oder asynchron), die Symmetrix-Hostanschlüsse (Kreuzverkabelung vorhanden = X-Link-Konfiguration, Kreuzverkabelung nicht vorhanden = U-Link-Konfiguration) und die Entfernung der beiden Standorte, sowie die Miteinbeziehung eines Globalspeichers.

Die nötigen Vorbereitungen, wie sie in den Kapiteln 3.1 bis 3.7 beschrieben sind, sind bei allen Konfigurationen zu treffen.

Statt synchronem SRDF ist stets auch alternativ semi-synchrones SRDF (s. 2.8.1) einsetzbar, falls der mögliche Verlust einer Schreib-IO pro (SRDF-)gespiegeltem Volume im K-Fall toleriert werden kann.

Weiterhin ist bei der Planung einer synchronen KS-Konfiguration genau abzuwägen, ob der Einsatz des Domino-Modus erforderlich ist. Die Entscheidungskriterien sind in Kap. 2.8.2 beschrieben. Die in Kap. 4.2 beschriebene Konfiguration mit GS als Schreibcache macht sogar den Einsatz des Domino Modus erforderlich.

4.1 Synchrone enge KS-Konfiguration (X-Link-Konf.)

Die folgende Abbildung zeigt das Prinzipbild der synchronen engen KS-Konfiguration. Sie besteht aus zwei Hosts (i.a. Produktions- und Standby-Host), die räumlich getrennt zusammen mit je einer Symmetrix betrieben werden. Die Hosts sind über eine Netzverbindung gekoppelt und haben jeweils Zugriff auf beide Symmetrix-Systeme. Jede der genannten Komponenten hat also eine Verbindung mit jeder anderen. Es handelt sich um eine Kreuzverkabelung, die einen Shared-Pubset-Verbund ermöglicht.

Grundsätzlich sollten alle Hardware-Ressourcen redundant ausgelegt sein, um Single Points of Failure auszuschließen, d.h. alle Platten sollten RAID1-geschützt sein, alle Verbindungen von den Hosts zu den Plattensteuerungen sowie die SRDF-Links sind wenigstens zwei-pfadig auszulegen und auf unterschiedlichen Wegen verlaufen und die Netzanbindungen zu den Anwendern sind redundant auszulegen. Ebenso sollte jede SRDF Source Unit wenigstens über zwei Remote-Link-Directoren Verbindung zu ihrer Target Unit haben. Dies stellt Hochverfügbarkeit weitgehend sicher. Für einen Einsatz von HIPLEX AF und vor allem für einen automatischen KS ist diese Redundanz zwingende Voraussetzung.

Die hier beschriebene Konfiguration wird im HIPLEX AF Manual [3] als „Grundkonfiguration“ bezeichnet.

Prinzipbild der synchronen engen KS-Konfiguration

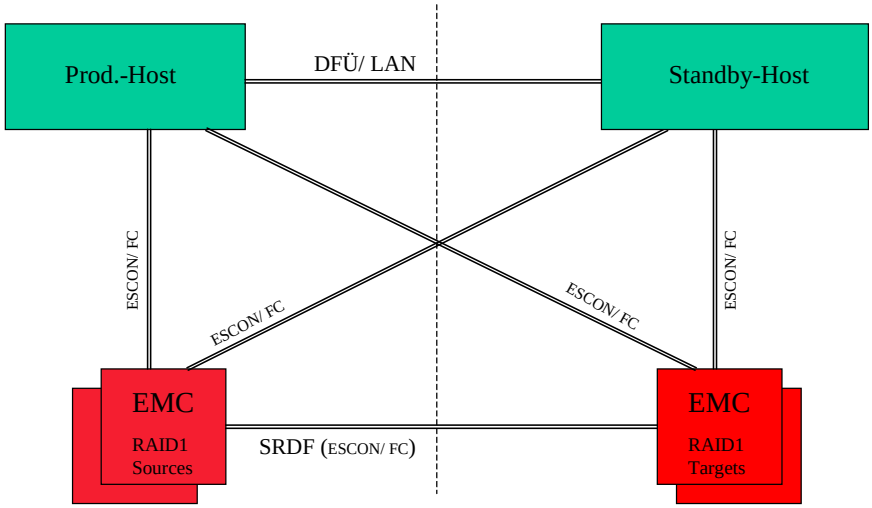


Abb. 4-1 Synchrone enge KS-Konfiguration

Von den zwei Hosts wird bei einer einseitigen KS-Konfiguration (2.3) einer als Produktiv- und einer als Standby-Host dienen, bzw. es liegt eine symmetrische KS-Konfiguration vor und beide Hosts spielen die Rollen von Produktiv- und Standby-Host gleichzeitig.

Die Konfiguration ist angebracht für Rechenzentren mit zwei separaten, mindestens durch eine Brandschutzmauer getrennten Räumen, die jeweils über eine eigene Stromversorgung verfügen oder besser noch für zwei RZs in unterschiedlichen Gebäuden, die als Produktions- bzw. Standby-RZ dienen können. Die Gebäude sind üblicherweise nicht weit voneinander entfernt und wenn kein öffentlicher Grund dazwischen liegt, spricht man üblicherweise von einer Campus-Lösung.

Falls die örtlichen Gegebenheiten zwischen den beiden Standorten jedoch die direkte Verlegung von (Glasfaser-)kabeln verhindern, muss man auf einen Netzanbieter ausweichen und evtl. „lose“ oder sogar asynchrone Konfigurationen (siehe 4.4) in Betracht ziehen. Eine Begrenzung der räumlichen Entfernung zwischen den RZs kann sich durch die Nutzung der Kreuzverbindungen ergeben: Will man bei Systemausfällen auch die Anwendungen auf dem Standby-Host mit Zugriff auf die Platten der Work-Symmetrix laufen lassen, dann muss berücksichtigt werden, dass die Performance des ESCON-Protokolls bei Entfernungen von mehr als 9 km merklich abnimmt. Prinzipiell kann so eine Konfiguration auch über größere Distanzen über eine gemietete Netzanbindung betrieben werden. Bei Nutzung von FC-Anschlusstechnik, möglich auf S1- und SX-Maschinen mit BS2000 OSD V5, bei der spezielle Flusskontrollmechanismen (Buffer Credits) genutzt werden können, kann die Distanz ohne Performanceeinbruch auf etwa 100 km erhöht werden (s. hierzu auch Kap. 3.5).

Die synchrone enge KS-Konfiguration ist also durch folgende Eigenschaften gekennzeichnet:
- Synchrone SRDF-Spiegelung
- Aktueller Datenbestand im Ausweich-RZ in einem K-Fall
- Kreuzverkabelung der spiegelbildlichen EMC-Steuerungen, Shared-Pubset-Verbund (MSCF)
- Angebracht bei Entfernungen zwischen Produktions- und Standby-RZ bis zu ca 10 km
- Ausfallzeit im K-Fall in der Größenordnung von ca. 30 min
- Symmetrische KS-Konfiguration möglich
- Optional Einsatz von HIPLEX AF mit automatischer Ausfallerkennung und automatischem Wiederanlauf möglich

Die X-Link-Konfiguration hat durch die Kreuzverkabelung den Vorteil, dass auf beiden Hosts mit beiden Plattensteuerungen gearbeitet werden kann (solange die Entfernung dieses zulässt). D.h. bei einem Systemausfall am Produktiv-Host kann jederzeit ein Switchover auf den Standby-Host (ohne Failover auf die Standby-Plattensteuerungen und zeitaufwendigen Failback) ausgeführt werden. Diese Möglichkeit kann ebenfalls zu Test- und Wartungszwecken genutzt werden. Weiterhin besteht die Möglichkeit, zu Wartungszwecken einen Failover auf die Work-Symmetrix auszuführen und weiter auf dem Work-Host - über die Kreuzverkabelung - zu arbeiten. Diese Konfiguration ist somit unter dem Gesichtspunkt der Hochverfügbarkeit den im Folgenden beschriebenen (losen) U-Link-Konfigurationen überlegen.

Bei Einsatz eines GS sind für ein Katastrophenschutz-Konzept zusätzliche Überlegungen erforderlich, die im nächsten Abschnitt dargestellt werden.

4.1.1 Einsatz von HIPLEX AF

HIPLEX AF kann zum einen für die Hochverfügbarkeit der Anwendungen genutzt werden, um bei Ausfall eines Systems oder wichtiger Anwendungsvoraussetzungen möglichst schnell auf den zweiten Host umzuschalten, zum anderen auch zum automatischen Katastrophenschutz, die Hochverfügbarkeit wird hier durch den KS nicht beeinträchtigt. HIPLEX AF kann bei dieser Konfiguration sinnvollerweise zwei Shared Pubsets nutzen, die zum einen als Überwachungs-Datenträger für die Watchdog-Funktionalität (s. 2.12) dienen, und auf denen die Daten der Switch-Units abgelegt werden (s. Kap. 2.13). Bei dieser Konfiguration liegen die besten Bedingungen für den automatischen KS mit HIPLEX AF vor.

Eine im Lieferumfang von HIPLEX AF enthaltene beispielhafte Generierungsprozedur für eine Switch-Unit mit Failover-Funktionalität geht daher auch von dieser Konfiguration aus. Diese Prozedur ist zugeschnitten auf den einfachsten Fall einer solchen Konfiguration mit einseitigem, asymmetrischem KS und Nutzung nur einer Switch-Unit (sogenannte Grundkonfiguration).

Der Ablauf des automatischen Failover mit HIPLEX AF ist in Kap. 5.1 beschrieben.

4.2 Synchrone enge KS-Konfiguration mit GS

Um einen Global Store in einer KS-Konfiguration zur Ablage von Schreibdaten kritischer Anwendungen (in DAB-GS-Caches oder GS-Volumes, s. Kap. 2.10) einsetzen zu können, ist zunächst ein Dual-GS (zwei GS-Units mit HW Duplication Feature und jeweils eigener Battery Unit) erforderlich. Das entspricht im Sinne einer Hochverfügbarkeits-Konfiguration den RAID1-Platten. Weiter müssen die beiden ausfallunabhängigen GS-Units (GSUs) räumlich getrennt werden. Produktions- und der Standby-Host müssen beide Zugriff zu jeweils beiden GS-Units haben, und die beiden BS2000-Systeme einen XCS-Verbund (und damit einen sogenannten Parallel Hiplex) bilden. Die Anbindung des GS über Glasfaserkabel erlaubt eine Entfernung von 70m [(1)], so dass diese Konfiguration lediglich eine Campus-Lösung sein kann.

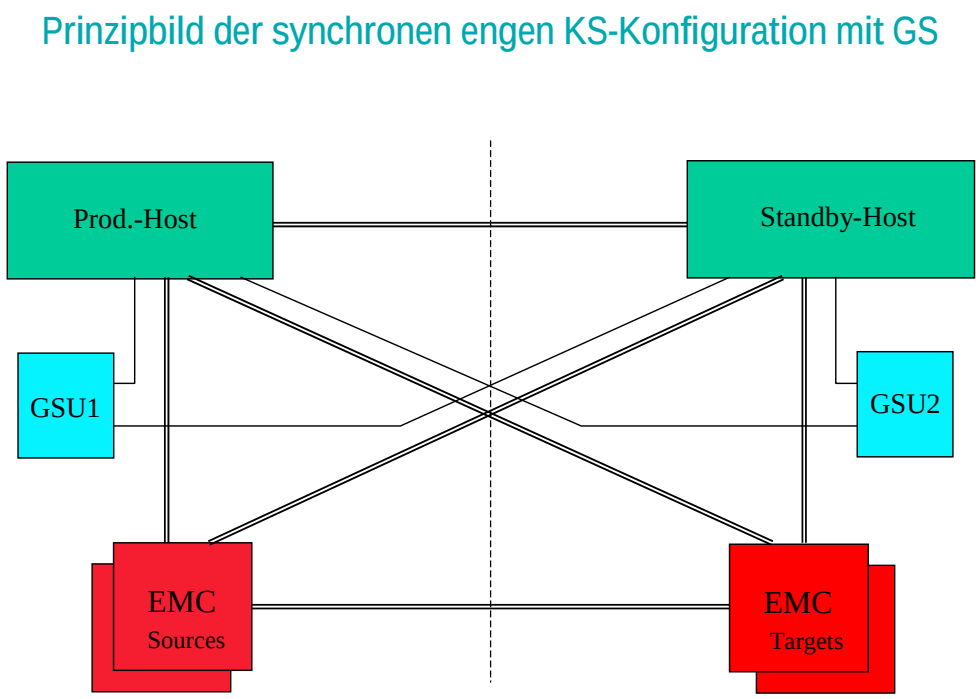


Abb. 4-2 Synchrone enge KS-Konfiguration mit GS

Gegenüber einer KS-Konfiguration ohne GS gibt es bei der Synchronen engen KS-Konfiguration mit GS nur dann neue Gesichtspunkte, wenn der Globalspeicher für geschäftskritische Daten genutzt wird:

- GS-Volumes (in Dual GS-Partitions) können nur dann für geschäftskritische Daten genutzt werden, wenn logisch voneinander abhängige Daten (wie Transaktion und Logging) nicht auf GS-Volumes einerseits und andere Speichersubsysteme andererseits verteilt sind, da diese keine Konsistenzgruppe (s. Kap. 2.11) bilden können.

- Werden Daten einer Anwendung, deren zugehörige Pubsets in einem Symmetrix Subsystem liegen, über DAB im GS schreibend zwischengepuffert, so müssen diese Pubsets im Symmetrix Subsystem und die DAB-Caches im GS zu einer Konsistenzgruppe zusammengefasst werden, um synchronen Katastrophenschutz für die Anwendung gewährleisten zu können (alle Daten im GS, die von einem DAB Schreib-Cache herrühren, befinden sich nur temporär im GS und werden, wenn sie von DAB asynchron auf die zugehörigen Symmetrix-Volumes geschrieben wurden, im GS wieder verdrängt).

  Die Notwendigkeit dieser Konsistenzgruppe erklärt sich folgendermaßen:
  Es gibt in dieser Konfiguration zwei denkbare Ausfallvarianten, bei denen Dateninkonsistenzen entstehen können (entweder wird auf Target Units weitergeschrieben, während die GSU2 veraltet, oder es wird auf GSU2 weitergeschrieben und die Target Units veralten):

  1. Es fällt zuerst die SRDF-Verbindung aus, während die Anwendungen noch weiterlaufen und danach fällt das gesamte Produktions-RZ aus. In diesem Fall können im Standby-RZ Daten-Inkonsistenzen entstehen, wenn weiterhin Daten aus den Globalspeicher-Caches von DAB auf die Source Units
     (aber wegen der ausgefallenen SRDF-Verbindung nicht mehr auf die Target Units) zurückgeschrieben werden.

[1] technisch möglich sind maximal 280m, eine solche Ausstattung muss im Einzelfall per Anfrage geklärt werden

2. Es fällt zunächst die Verbindung des Produktions-Hosts zur GSU2 aus, während die Anwendungen noch weiterlaufen, und danach fällt das gesamte Produktions-RZ aus (Abb. 4-2). Dadurch wären die Daten auf den Target Units u.U. aktueller als die Daten in der GSU2.

Die geschilderten Überlegungen zeigen, dass in einem Parallel HIPLEX bei Einsatz von DAB-Schreibcaches im GS für Dateien auf einem Symmetrix SRDF-Verbund die Eigenschaften einer Konsistenzgruppe für die Anwendungsdaten nicht notwendigerweise vorliegen.

Für den oben beschriebenen Fall 1 gibt es den Domino-Modus in der Symmetrix, dieser bewirkt, dass IO‘s auf eine SRDF Unit nicht mehr zugelassen werden, sobald eine der drei Komponenten Source, Target oder der SRDF-Link ausfallen.

Um Inkonsistenzen auch in dem oben genannten Fall 2 zu verhindern und eine Konsistenzgruppe zu etablieren, benötigt man auch für die DAB-Caches den Domino-Modus im GS (s. 2.8.2), näheres zum DAB-Domino-Modus findet man im DAB-Handbuch [2]. Der Domino-Modus im DAB bewirkt, dass bei Ausfall einer GS-Unit eines dualen GS – und damit auch bei einem Verbindungsausfall - alle folgenden Schreib IO’s fehlerhaft beendet werden und somit die Anwendung wie auch bei Fall 1 abbricht oder stehenbleibt. Wenn die Fehlerursache geklärt ist und ein K-Fall ausgeschlossen werden kann, kann der Domino-Modus deaktiviert und temporär mit nur einer GSU weitergearbeitet werden.
Während der SRDF-Domino-Modus pro logischem Volume eingestellt werden kann, wirkt der DAB-Domino-Modus global auf die Gesamtheit aller DAB-Caches im Globalspeicher.

Für einen rundum sicheren KS in einem Parallel HIPLEX, in dem der GS als DAB-Schreib-Cache genutzt wird, ist demnach sowohl der Domino-Modus für SRDF als auch der Domino-Modus für DAB-Caches im GS zu verwenden. Für GS-Volumes steht z.Z. kein Domino-Modus zur Verfügung.

Die synchrone enge KS-Konfiguration mit Nutzung des GS für geschäftskritische Daten ist also durch folgende Eigenschaften gekennzeichnet:

- Synchrone SRDF-Spiegelung mit Domino-Modus
- Aktueller Datenbestand im Ausweich-RZ in einem K-Fall
- Kreuzverkabelung der spiegelbildlichen EMC-Steuerungen, Shared-Pubset-Verbund (MSCF)
- „Geringe“ räumliche Entfernung zwischen Produktions- und Standby-RZ (bis zu 70m)
- Ausfallzeit nach Erkennung des K-Falls in der Größenordnung von ca. 45 min
- Symmetrische KS-Konfiguration möglich
- Optional Einsatz von HIPLEX AF mit automatischer Ausfallerkennung und halbautomatischem Wiederanlauf möglich
- Nutzung von DAB-Schreibcaching im GS mit Domino-Modus möglich.

Durch den Domino-Modus muss bei Ausfall der SRDF-Verbindungen und auch bei Ausfall der Verbindung zwischen Work-Host und einer GSU eine Unterbrechung der Anwendung in Kauf genommen werden. Gleiches gilt auch bei Ausfall der Standby Symmetrix oder der entfernten GSU. In diesen Fällen muss der Domino-Modus für SRDF oder DAB deaktiviert und ggf. die Anwendungen neu gestartet werden. Kann der Fehler nicht in kurzer Zeit behoben werden, muss man ohne Domino-Modus (bzw. je nach Art des Ausfalls ohne SRDF-Spiegelung oder ohne Dual-GS) und mit der oben genannten Gefahr von Inkonsistenzen oder Datenverlust bei einem möglichen K-Fall in dieser Phase weiterarbeiten. Dies wird man nur dann tun, wenn der aufgetretene Fehler lokalisiert ist, und ausgeschlossen werden kann, dass er im Rahmen einer sich entwickelnden Katastrophe auftrat.

Der Domino-Modus schließt einerseits das Risiko von Datenverlust und Inkonsistenzen auch bei den genannten Rolling Disaster Fällen aus, andererseits wird eine Konsistenzgruppe von Platten zusammen mit GS-Caches geschaffen, deren Ausfallwahrscheinlichkeit wiederum höher ist als die eines SRDF-Paares ohne Domino und GS-Cache, so dass eine Abschwächung der Hochverfügbarkeit in Kauf genommen werden muss.

Abzuwägen bei der Entscheidung für oder gegen den Einsatz des Domino-Modus sind die folgenden grundsätzlichen Aussagen:

1. Nahezu jeder Einzel-Ausfall der relevanten Daten-Behälter Symmetrix und GS führt zum Abbruch der Anwendungen – die Verfügbarkeit ist damit beeinträchtigt. Da in der Folge, i.d.R. abhängig von der Analyse eines Technikers, eine Umschaltung auf den Standby-Rechner durchgeführt werden soll, hängt der Grad der Beeinträchtigung der Verfügbarkeit von der Zeitdauer der Techniker-Analyse und der Zeitdauer der Umschaltung ab.

2. Dafür gilt die höchste Sicherheitsstufe bei maximaler Performance: kein denkbarer Ausfall kann zu einem inkonsistenten Datenbestand auf der Standby-Seite führen und IO-Engpässe werden vom GS abgefangen.

Falls die in Punkt 1 genannte Beeinträchtigung der Hochverfügbarkeit nicht akzeptabel ist, muss auf das Schreibcaching im GS für geschäftskritische Daten verzichtet werden.

4.2.1 Nutzung und Arbeitsweise von HIPLEX AF

Für den Einsatz zur Katastrophenvorsorge wird der Domino-Modus für die SRDF-Spiegelung sowie für den DAB-Einsatz im GS vorausgesetzt. Für einen automatischen Wiederanlauf durch HIPLEX AF sind zusätzliche Schritte in der Switch-Unit zu implementieren: so muss zunächst am Standby-Host der DAB-Domino-Modus ausgeschaltet werden, um in jedem Fall auch mit nur einer GS-Unit weiterarbeiten zu können, und dann muss wegen der Umschaltung der DAB-Caches vom Produktions- auf den Standby-Host beim Importieren der von DAB im GS zwischengepufferten Datenpubsets jeweils die Meldung EGC0502 des GS-Managers automatisch beantwortet werden.

Ein automatischer Wiederanlauf mit Failover auf die Standby-Symmetrix durch HIPLEX AF ist aus folgenden Gründen nicht unkritisch: bei der Nutzung des Domino-Modus führt auch jeweils der Ausfall der Standby Symmetrix, der Ausfall aller SRDF-Links, der Ausfall einer der beiden GS-Units oder einer Verbindung eines Hosts zu einer GS-Unit jeweils zu einem Ausfall der Anwendungen. Es ist Sache der Systemadministration oder eines Service-Technikers zu entscheiden, ob der Ausfall begrenzt ist und keinen Ausfall des gesamten RZs (oder auch nur des Produktions-Hosts oder eines Symmetrix Systems am Produktions-Standort) zur Folge haben wird. Wird auf begrenzten Fehler entschieden und deshalb der Domino-Modus von der Systemadministration per Kommando ausgeschaltet, so würde bei einer Fehlentscheidung und nachfolgendem K-Fall mit automatischer Umschaltung durch HIPLEX AF ein vermutlich inkonsistenter Datenbestand am Standby-RZ aktiviert.

Diese etwas kritische Stelle kann aber dadurch entschärft werden, dass im Rahmen der Entscheidung, ob nach dem ersten Ausfall weitergearbeitet werden soll, eine Prozedur gestartet wird, die den Domino-Modus ausschaltet und weiter dafür sorgt, dass bei nachfolgenden Fehlerfällen kein automatischer Failover auf die SRDF Target Units durch HIPLEX AF gestartet wird.

Bei Einsatz des Domino-Modus kann also ein halbautomatisches KS-Konzept gewählt und die Überwachung der Switch-Units derart erweitert werden, dass jeweils bei Ausfall aller SRDF-Links, bei Ausfall einer der Symmetrix Subsysteme oder bei Ausfall einer GS-Unit ein Techniker und/ oder Systemverwalter benachrichtigt wird, der dann zunächst die Situation überprüft, bevor entschieden wird, ob die Anwendung ohne Domino weiterlaufen soll, da der Fehler erkannt und behebbar ist, oder ob ein Failover auszuführen ist, da ein K-Fall vorliegt. Mit etwas Zusatzaufwand ist, wie oben angedeutet, sogar ein vollautomatisches KS-Konzept mit HIPLEX AF Mitteln möglich.

Bei der Prozedur für den Failback ist für Testeinsätze ein Parameter zu setzen, der den Domino-Modus vor der Wiederinbetriebnahme der SRDF Source Units deaktiviert und am Ende des Failbacks wieder aktiviert, da durch den Domino-Modus bei einem Ausfall die Source Units „disabled“ werden (s.a. Kap. 2.8.2). Bei einem wirklichen Ausfall der Symmetrix wird dies i.d.R. ein Techniker vornehmen. Ebenso kann für den DAB-Domino-Modus ein entsprechendes Kommando in die Switch-Unit eingebaut werden.

4.2.2 Besonderheiten bei Failover und Failback

Der Failover erfolgt wie im Kap. 5 beschrieben. Sobald die Target Units in der Standby Symmetrix freigeschaltet sind, können die Datenpubsets importiert werden und im Rahmen des Imports werden die noch zu sichernden Daten in der GSU2 von DAB automatisch auf die Target Units geschrieben. Für Privatplatten, die mit GS gecached werden, werden noch ausstehende Auslagerungen bei der ersten Belegung einer Platte ausgeführt. Aus Sicht des Caching sind hier also keine Zusatzmaßnahmen erforderlich.

Falls mit Domino-Modus gearbeitet wird, muss dieser im DAB deaktiviert werden, bevor die Target Units belegt werden. Der Domino-Modus der Symmetrix muss beim Failover nur dann abgeschaltet werden, wenn die Work-Symmetrix noch läuft, z.B. bei einer Umschaltung aus Testzwecken.

4.3 Synchrone lose KS-Konfiguration (U-Link-Konf.)

Die synchrone lose Konfiguration besteht, wie auch die synchrone enge, aus einem Work- und einem Standby-RZ, die so ausgelegt sind, dass jedes bei Komplettausfall des anderen auch die Gesamtheit der Anwendungen beider RZs übernehmen kann, d.h. die Rechenleistung, der Plattenspeicherbedarf, die Netzanbindung sind annähernd spiegelbildlich. Auch die Redundanz der Hardware-Ressourcen sollte, wie in Kap. 4.1 beschrieben, gegeben sein. Mit dem Begriff „lose“ ist hier gemeint, dass es keine Kreuzverbindungen zwischen Produktiv-Host und Standby-Symmetrix sowie zwischen Standby-Host und Work-Symmetrix (und damit insbesondere keinen Shared-Pubset-Verbund) gibt. Der Verzicht auf diese Kreuzverkabelung kann dann angebracht sein, wenn es sich um größere Entfernungen handelt, bei denen die standortübergreifenden Anwendungs-IOs, im Gegensatz zu den SRDF-Transfers, nicht mit der nötigen Performance durchgeführt werden könnten.

Durch das Fehlen der Kreuzverbindungen entfallen hier die Shared Pubsets für HIPLEX AF und die gegenseitige Überwachung der Systeme kann lediglich über die Kommunikation über die Netz-Verbindungen der Hosts (BCAM/ MSCF) erfolgen. Die Möglichkeit, von beiden Hosts mit beiden Symmetrix-Subsystemen arbeiten zu können, entfällt, und somit sind die Hochverfügbarkeits-Eigenschaften gegenüber der einer engen Konfiguration leicht abgeschwächt.

Zusammengefasst lässt sich die Synchrone lose KS-Konfiguration kennzeichnen durch die folgenden Eigenschaften:

- Synchrone SRDF-Spiegelung
- Aktueller Datenbestand im Ausweich-RZ in einem K-Fall
- Angebracht für Entfernungen zwischen Produktions- und Standby-RZ von 10 km bis 200 km
- Ausfallzeit im K-Fall in der Größenordnung von ca. 30 min
- WDM-Technologie (oder WAN-Verbindung) für SRDF-Verbindungen
- Einsatz von HIPLEX AF mit (halb)automatischer Ausfallbehandlung möglich, hierfür sind zwei getrennte, unabhängige BCAM Verbindungen nötig (HIPLEX MSCF V3.0)
- Konfiguration kann symmetrisch betrieben werden
- Voraussetzung: Hotspots der Anwendungen sind auch mit der SRDF-Zeitverzögerung noch ablauffähig (s. 3.4.3).

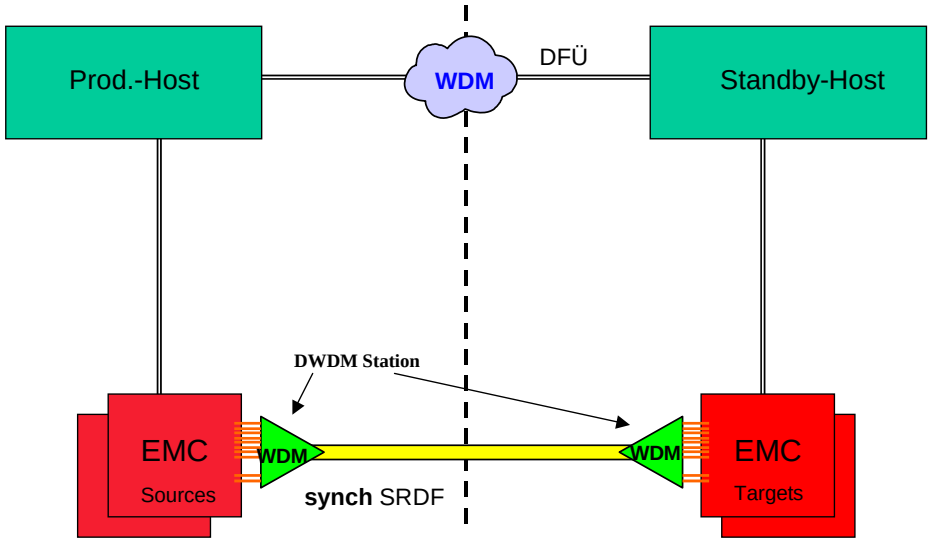


Abb. 4-3 Synchrone lose KS-Konfiguration

4.3.1 Nutzung und Arbeitsweise von HIPLEX AF

Da in dieser Konfiguration keine Shared Pubsets möglich sind, hat nur das HIPLEX AF Monitoring am Work-System Zugriff auf die Daten der Switch-Unit für die Anwendungen. Daher wird hier zusätzlich eine Hilfs-Switch-Unit am Standby-System definiert. Diese überwacht lediglich den Work-Host und das Symmetrix-Subsystem und führt im K-Fall den Failover für die Standby Symmetrix durch. Nachdem die Target Units freigeschaltet sind, stehen die Daten der produktiven Switch-Unit zur Verfügung und diese kann gestartet werden. Die Überwachung der Systeme erfolgt hier nur über die MSCF-Verbindungen und nicht über die Watchdog-Funktion mit Shared Pubsets (s. 2.12). Für die automatisierte Überwachung müssen also mindestens zwei redundante MSCF/ BCAM Verbindungen vorhanden und HIPLEX MSCF 3.0 im Einsatz sein. Welche Kriterien einen (halb-)automatischen Failover auslösen sollen, bestimmt letztlich der Systembetreiber. Die Definition der Switch-Units für die Anwendungen sind auf einem Daten- oder dem Systempubset in der Work-Symmetrix abgelegt, welches, wie die Anwendungsdaten, per SRDF gespiegelt wird.

Durch das Fehlen der Kreuzverkabelung ist die Möglichkeit des Switchover ohne Failover (also ein Umschalten auf den Standby-Host ohne Umschaltung auf die Standby-Symmetrix, s.a. Abb. 5-2) nicht gegeben. Ein automatisches Umschalten bei einem Systemausfall (nicht K-Fall) bedeutet somit stets auch ein Freischalten der Target Units, also einen kompletten Failover (und späteren Failback). Da dies in den meisten Fällen mehr Aufwand darstellen wird, als der Neustart des Systems, empfehlen wir deshalb, in losen Konfigurationen mit HIPLEX AF einer Halb-Automatik den Vorzug zu geben, d.h. die Switch-Unit so zu programmieren, dass ein Systemadministrator benachrichtigt und Failover und Switchover erst auf Nachfrage ausgeführt werden. Trotzdem ist natürlich auch hier ein vollautomatisierter KS möglich.

Auch die Möglichkeit des Switchover ohne Failover zu Testzwecken ist in der losen Konfiguration wegen der fehlenden Kreuzverkabelung nicht gegeben.

4.3.2 Besonderheiten bei Failover und Failback

Es gibt keinen Shared-Pubset-Verbund und deshalb keinerlei Aktionen im Zusammenhang mit Shared Pubsets (z.B. kein Einrichten eines Shared-Pubset-Verbunds beim manuellen Failback).

Beim Einsatz von HIPLEX AF V3 zur automatischen Ausfallerkennung und Ausfallbehandlung wird, wie im vorigen Kapitel beschrieben, das Monitoring einer Hilfs-Switch-Unit auf dem Standby-Host den Ausfall des Produktiv-Hosts und den Ausfall der Work-Symmetrix erkennen und den Failover auf die Standby-Symmetrix ausführen. Danach startet diese die produktive Switch-Unit, deren Daten nach dem Failover zugreifbar sind.

Der Failback kann mittels HIPLEX AF automatisiert werden, s.a. Kap. 6, auch er kann durch eine Hilfs-Switch-Unit am Work-System gestartet werden. Zusammenfassend sei gesagt, dass Failover und Failback mit HIPLEX AF bei dieser Konfiguration ähnlich verlaufen, wie bei der „synchronen engen Konfiguration“, jedoch sind die Vorbereitungen hierfür etwas umfangreicher.

4.4 Asynchrone lose KS-Konfiguration

Selbstverständlich gibt es viele verschiedene Varianten von asynchronen KS-Konfigurationen, da ihr Charakteristikum ja nur darin besteht, dass die in einem K-Fall am Ausweich-RZ zur Verfügung stehenden Datenbestände zwar stets konsistent, aber in der Regel nicht aktuell sind. So fällt insbesondere auch die Variante des regelmäßigen Bandtransports (s. 3.4) in die Kategorie der asynchronen KS-Konfigurationen.

Die hier vorgestellte asynchrone lose Konfiguration entspricht weitgehend der synchronen losen Konfiguration aus Abb. 4-4. Das synchrone SRDF wird dabei durch ein asynchrones SRDF ersetzt, den Adaptive Copy Mode (s.a. Kap.2.8.1), wodurch man die möglichen negativen Performance-Auswirkungen (s. Kap. 3.4.3) einer synchronen SRDF-Spiegelung auf kritische Hotspot-Anwendungen vermeiden kann.

Damit ist diese funktional schwächere Konfiguration gegenüber der synchronen losen Konfiguration nur dann zu bevorzugen, wenn dies aus Performance-Gesichtspunkten (z.B. wegen sehr großer Entfernung zwischen den beiden Standorten) erforderlich ist. Durch den asynchronen SRDF-Betrieb kann die WDM Kopplung ggf. auch durch eine weniger breitbandige ATM Verbindung ersetzt werden.

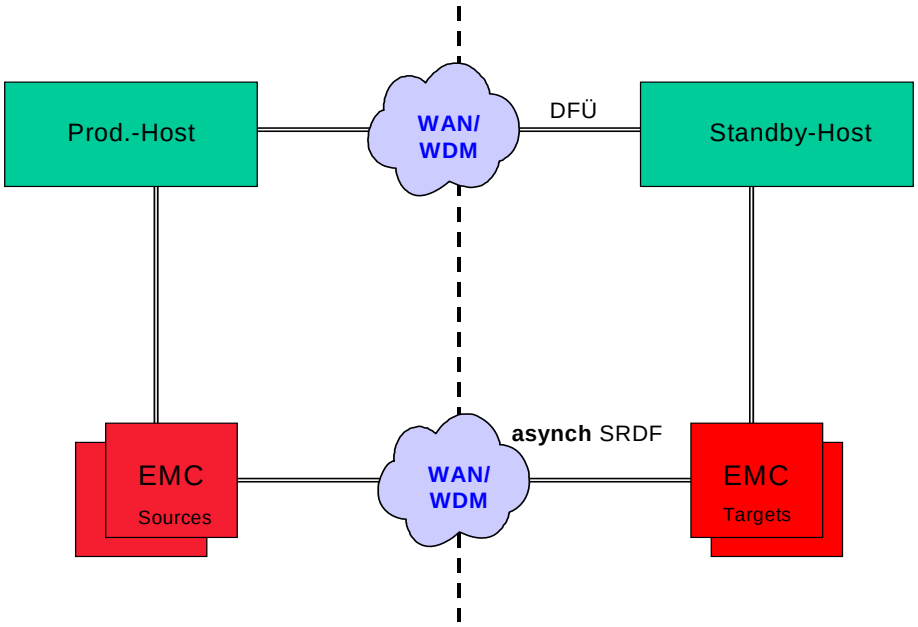


Abb. 4-4 Asynchrone lose KS-Konfiguration

Weil der Datenabgleich bei asynchronem SRDF-Betrieb nicht kontinuierlich und nicht in der richtigen Reihenfolge der Schreib-IOs stattfindet, müssen für konsistente Pubset-Daten auf den Target Units die Anwendungen regelmäßig angehalten, bzw. Datenbanken (Oracle) in den „Backup-Mode“ geschaltet werden. Nachdem in dieser Anwendungspause die SRDF-Targets abgeglichen sind, sie also keine invalid tracks mehr haben, kann man diesen synchronisierten Konsistenzstand auf einfache Weise einfrieren, wenn man BCV-Volumes für die Target Units einsetzt. Nach der Abtrennung der BCV-Spiegelvolumes von den Target Units im Ausweich-RZ kann die Anwendung wieder gestartet werden. Die Dauer ein solchen Unterbrechung wird nur im Minutenbereich liegen.

Die folgende Abb. 4-5 zeigt eine solche Lösung mit Multi-BCVs und deren Ablaufreihenfolge beim periodischen Schreiben eines Konsistenzpunktes:
Es werden hier zwei sogenannte Multi-BCVs pro Target Unit genutzt: Die BCVs 1 und 2 enthalten alternierend eine den aktuellen Stand der Target Unit und eine den letzten eingefrorenen Stand. Der Wechsel zwischen BCV 1 und BCV 2 besteht jeweils aus einem Differenzabgleich (Symmetrix-Funktion „Multi-BCVs“) und benötigt daher wenig Zeit. Zum regelmäßigen Erstellen eines konsistenten Datenbestandes müssen die Anwendungen pausieren, bis Target Units und Source Units synchronisiert und das aktuelle BCV von der Target Unit getrennt ist. Danach wird die Spiegelung der Target Unit mit dem zweiten BCV wieder aufgenommen und die Anwendung kann weiter laufen.

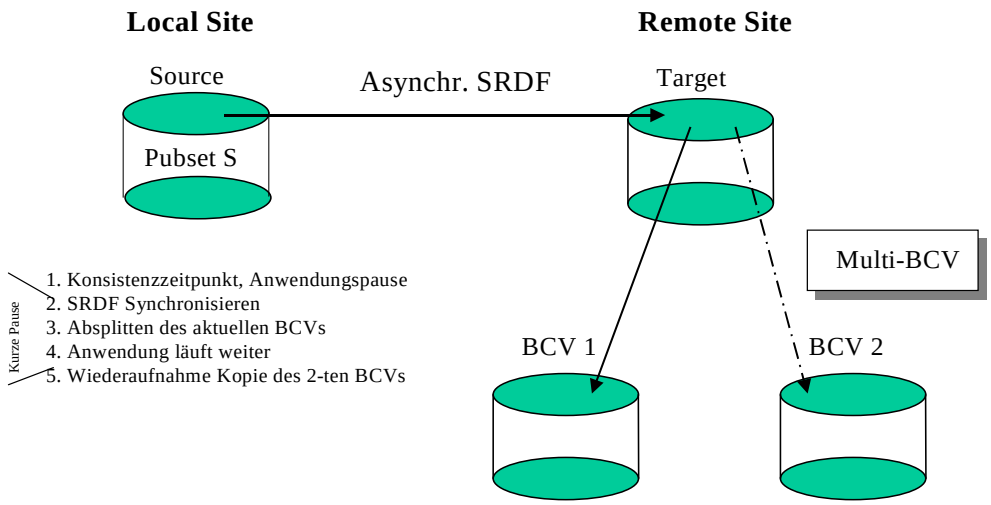


Abb. 4-5 Asynchrones SRDF mit Multi-BCVs

Bei dieser Konfiguration besteht somit ein zusätzlicher (zusätzlich zu SRDF-Targets und RAID1-Spiegeln) Plattenaufwand von zwei BCVs pro Nutzplatte, das Arbeiten mit zwei BCVs hat aber den Vorteil, dass stets der einzufrierende Datenbestand schon auf einem der BCVs bereitsteht. Will man auf ein zweites BCV verzichten, benötigt man ein größeres Zeitfenster, da das BCV bei jedem Abgleich erst zugeschaltet, synchronisiert und wieder weggeschaltet werden muss.

Diese Aynchrone lose KS-Konfiguration ist gekennzeichnet durch die folgenden Eigenschaften:
- Asynchrone SRDF-Spiegelung (Adaptive Copy Mode)
- Datenbestand des Zeitpunkts des letzten BCV-„Einfrierens“ im Ausweich-RZ in einem K-Fall
- Größte räumliche Entfernungen ohne merklichen Performanceverlust
- Ausfallzeit im K-Fall in der Größenordnung von ca. 30-45 min
- Einsatz von HIPLEX AF mit (halb)automatischer Ausfallbehandlung möglich, hierfür sind zwei getrennte unabhängige BCAM Verbindungen erforderlich
- Konfiguration kann symmetrisch betrieben werden
- Zusätzlicher Plattenbedarf durch TimeFinder/Mirror Spiegelplatten.

Während die beschriebene Lösung bei EMC unter dem Begriff Data Mobility (oder SRDF/DM) firmiert, ist mit der Variante SRDF/A von EMC eine weitere asynchrone Konfigurations-Möglichkeit am Markt – diese setzt Symmetrix-Modelle vom Typ DMX voraus, besitzt die Vorteile der weniger breitbandigen und dafür preisgünstigeren asynchronen Konfigurationen und liefert im K-Fall weit aktuellere Daten als die oben beschriebene SRDF/DM-Variante.
Für die Implementierung einer asynchronen SRDF/A KS-Konfiguration im BS2000-Umfeld wenden Sie sich bitte an Ihr Competence-Center von Fujitsu Technology Solutions.

4.4.1 Nutzung und Arbeitsweise von HIPLEX AF

Die Arbeitsweise von HIPLEX AF ist ähnlich wie bei der synchronen losen KS-Konfiguration. Wir empfehlen aber auch hier, auf einen automatisierten KS zu verzichten. Zu dem in Kap. 4.3.1 genannten Grund kommt hier noch die Tatsache, dass aufgrund der Asynchronität jede Umschaltung mit der Reaktivierung eines eingefrorenen Datenstandes am Standby-RZ und mit dem Verwurf der möglicherweise noch lesbaren aktuellen Daten am Produktiv-RZ verbunden ist. Jede Umschaltung bedeutet also Datenverlust. Deshalb sollte für diesen schwerwiegenden Eingriff einer Umschaltung keine automatische Ausfallerkennung verantwortlich sein. Bei einer asynchronen Konfiguration ist es daher immer angebracht, dass ein qualifizierter Mitarbeiter die Situation beurteilt, bevor ein Failover ausgeführt wird. HIPLEX AF kann natürlich als Überwachungsinstanz und für den „Failover auf Knopfdruck“ eingesetzt werden, von einer automatischen Umschaltung im K-Fall sollte man jedoch Abstand nehmen.

4.4.2 Besonderheiten bei Failover und Failback

Beim Failover werden, wie auch in Kapitel 5, Punkt 5 beschrieben, die SRDF Target Units freigeschaltet. Hierbei stellt man fest, ob noch Schreibaufträge für Target Units offen sind (invalid tracks) oder nicht. Falls ja, was beim Produktivbetrieb anzunehmen ist, dann muss zunächst die BCV-Spiegelung für die Target Units beendet werden. Im nächsten Schritt werden zunächst die (inkonsistenten) Target Units mit Hilfe der zum vorherigen Konsistenzzeitpunkt gesicherten BCVs rekonstruiert. Erst danach geht es, wie in Kapitel 5 beschrieben, weiter. Diese Zusatzschritte können auch beim Einsatz von HIPLEX AF in Prozedurform implementiert und von der Switch-Unit gestartet werden.

Der Failover kann bei dieser Konfiguration beschleunigt werden, indem mit den BCVs, die den letzten konsistenten Sicherungsstand enthalten, weitergearbeitet wird und die Target Units vorläufig gar nicht zum Einsatz kommen. Dafür werden die BCVs zugeschaltet und dann die Datenpubsets am Standby-Host importiert. Damit bleiben jedoch die Target Units sowie die zuletzt zugeschalteten BCVs ungenutzt. Zu gegebenem Zeitpunkt, spätestens jedoch vor dem Failback, muss daher der Betrieb unterbrochen und auf die Target Units umgeschaltet werden, die zuvor mit den Daten der aktuellen BCVs abgeglichen worden sind.

4.5 Kombinierte KS-Konfigurationen

Soll eine bereits vorhandene lokale Hochverfügbarkeits-Konfiguration zu einer KS-Konfiguration erweitert werden, besteht die Möglichkeit, einen zusätzlichen Host und ein zusätzliches Symmetrix System an einem weiter entfernten Standort, der lediglich die Aufgabe eines Katastrophenschutzes wahrnimmt, an diese Konfiguration zu koppeln. Die zusätzliche Spiegelung der Nutzdaten zu der dritten Symmetrix kann dann durch das Zwischenschalten von BCVs realisiert werden. Wir sprechen dann von einer kaskadierten KS-Konfiguration. Abb. 4-6 zeigt eine solche Konfiguration, die Spiegelung der Nutzdaten ist anhand eines Volumes mit den zugehörigen SRDF Target Units und BCVs dargestellt.

Die vorgestellte Variante stellt eine asynchrone KS-Konfiguration dar, weil die Synchronisation der Daten zur entfernten Target Unit erst bei Abtrennen des lokalen BCVs durchgeführt wird. Also liegt mit dieser Konfiguration eine Kaskadierung einer synchronen engen X-Link-Konfiguration mit einer asynchronen losen U-Link-Konfiguration vor.

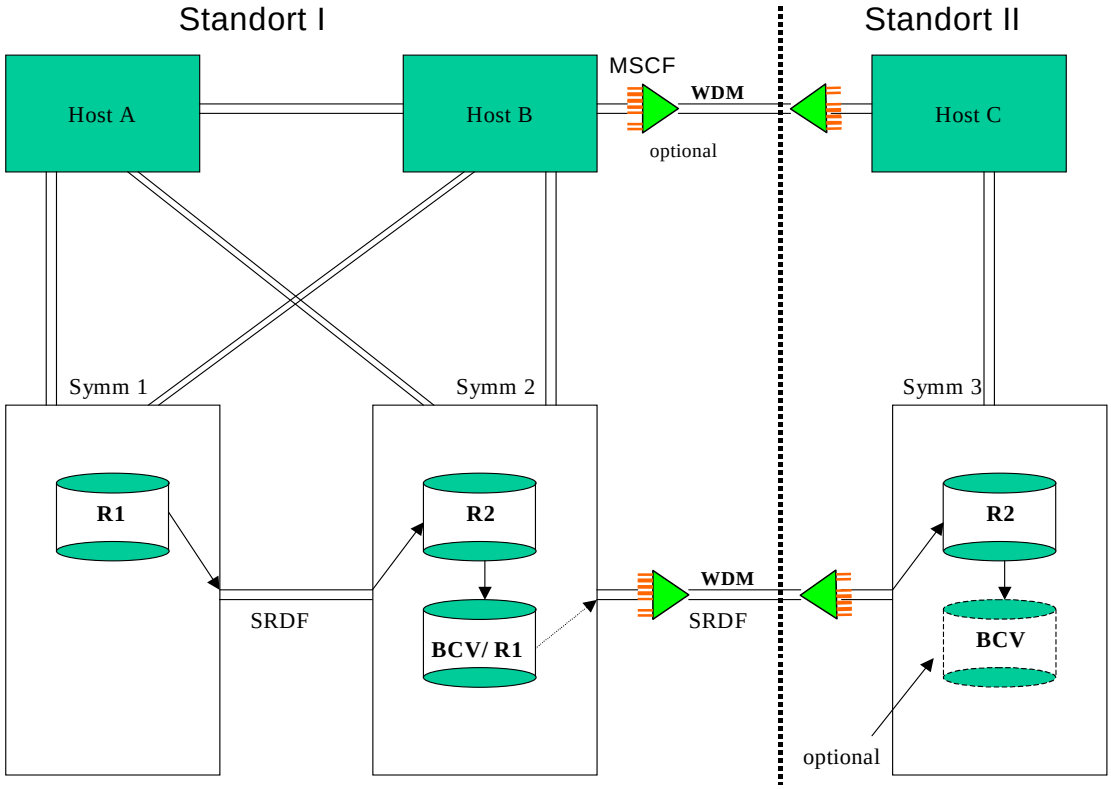


Abb. 4-6 Kombinierte KS-Konfiguration

Jede SRDF-Target Unit wird zusätzlich durch ein BCV gespiegelt, von dem wiederum zu festgelegten Zeiten eine SRDF-Synchronisierung auf ein entferntes Target in der Symmetrix 3 vorgenommen wird. In einem K-Fall am Standort I besteht dann die Möglichkeit, den Produktivbetrieb auf Host C mit der Symmetrix 3 wieder aufzunehmen, deren Daten dann vom letzten Synchronisationszeitpunkt sind.

Um sich bei diesem Verfahren auch für den Fall abzusichern, dass **während** der Synchronisierung von Symmetrix 2 nach Symmetrix 3 die Symmetrix-Systeme am Standort 1 ausfallen, und somit nicht konsistente Daten auf Symmetrix 3 vorliegen, können die Targets in Symmetrix 3 zusätzlich durch BCVs gesichert werden, die vor jeder Synchronisierung von den Target Units abgetrennt und danach wieder zugeschaltet werden. Wird auf die BCVs in Symmetrix 3 verzichtet, muss in so einem Fall eine Bandsicherung verfügbar sein.

Die Vorteile einer solchen Konfigurations-Variante sind

- Ausgeprägte HV-Konfiguration, die z.B. während der Produktion Wartungsarbeiten an einem abgeschalteten Symmetrix System erlaubt, ohne die Produktion dabei auf Betriebsmittel eines entfernten Ausweich-RZ umzuschalten.
- Kein Performance-Verlust durch synchrones SRDF über weite Entfernung, die Spiegelung ins entfernte Standby-RZ geschieht asynchron.

4.6 Konfigurationen mit mehr als einem Symmetrix Subsystem

Sind an einem Produktions-Host mehr als ein Symmetrix Subsystem angeschlossen, so ist die Aufteilung der Daten auf die verschiedenen Subsysteme sehr sorgfältig zu planen, um ein bündiges KS-Konzept zu erhalten.

Die Problempunkte einer solchen Konfiguration hinsichtlich Katastrophenschutz liegen auf der Hand:
Im Rahmen eines Rolling Disasters können die Symmetrix Subsysteme zu verschiedenen Zeitpunkten ausfallen. Wenn logisch abhängige Daten auf zwei Symmetrix Subsysteme verteilt sind, so kann ein derartiges Ausfallszenario zu inkonsistenten Datenständen führen – nämlich z.B. dann, wenn zunächst die SRDF-Verbindungen des einen und zeitlich später die SRDF-Verbindungen des anderen Symmetrix Subsystems zum jeweiligen Target-System ausfallen, kein Domino-Modus eingeschaltet ist und somit auf die Source Units weiter geschrieben wird, während im Standby-RZ die gespiegelten Symmetrix Subsysteme zu verschiedenen Zeitpunkten keine Schreib-IOs des Produktions-RZ mehr entgegennehmen können. So könnte für eine Datenbank z.B. im Standby-RZ der Gleichstand von ausgeführten Transaktionen und ausgeführtem Logging gefährdet sein.

Um dieses Risiko auszuschließen, kann entweder mit dem Domino-Modus (für alle SRDF-Paare mit kritischen Daten) gearbeitet werden, oder alternativ (und die Hochverfügbarkeit nicht beeinträchtigend) durch administrative Maßnahmen dafür gesorgt werden, dass alle Daten einer Anwendung gemeinsam in einem Symmetrix Subsystem abgelegt werden, also das Konzept einer Konsistenzgruppe (siehe Kap. 2.11) angewendet wird.

Auch eine automatische Anwendungsumschaltung via HIPLEX AF ist weniger komplex bei sauber getrennten Symmetrix Subsystemen, wenn es darum geht, auf den Ausfall lediglich eines Symmetrix Subsystems zu reagieren. Prinzipiell jedoch sind mit HIPLEX AF auch enge KS-Konfigurationen mit mehreren Symmetrix Subsystemen pro Standort beherrschbar.

4.7 Trennung von Anwendungen durch VM2000

Zum Abschluss dieses Kapitels wollen wir noch auf die Vorteile hinweisen, die der Einsatz von VM2000 auch im Hinblick auf KS haben kann – unabhängig von der Art der hier besprochenen Konfigurationen. Da unterschiedliche Anwendungen oft auch mit unterschiedlichen Systemvoraussetzungen arbeiten, kann es sehr nützlich sein, jeder Produktivanwendung ein eigenes Gastsystem, also eine VM zur Verfügung zu stellen. Für die K-Fall Vorsorge wird dann spiegelbildlich ebenfalls pro Anwendung eine „Spiegel-VM“ im Standby-RZ eingerichtet und mit den nötigen Werten für CPU- und Arbeitsspeicherbedarf versorgt. Eine Anwendung, die somit in beiden RZs auf einer eigenen VM ablauffähig ist, kann, wie auch die Switch-Unit bei HIPLEX AF, als Umschalteinheit bezeichnet werden, auch wenn HIPLEX AF nicht im Einsatz ist. Das Gastsystem kann dann fortwährend laufen, oder es wird erst in einem K-Fall aktiviert, s.a. Kapitel 3.6.1. Hierdurch wird eine bessere Übersichtlichkeit bei einem Failover erreicht. Sobald der Host im zweiten RZ auch produktiv genutzt wird und nicht nur als Testanlage oder reiner Standby-Host, empfehlen wir diese Vorgehensweise, um Anwendungen logisch voneinander zu trennen, die u.U. mit völlig unterschiedlichen Systemparametern und einer anderen Geräteausstattung arbeiten.

Die Abb. 4-7 zeigt als Beispiel die Aufteilung dreier Anwendungen „Finanzen“, „Personal“ und „Lager“ in einer symmetrischen KS-Konfiguration, von denen die ersten beiden im Normalbetrieb im Work- und die dritte im Standby-RZ laufen (aus Sicht der Anwendung „Lager“ sind die Begriffe Work- und Standby-RZ dann vertauscht). Die grau eingefärbten VMs sind im Normalbetrieb ungenutzt oder dienen nur als Testsysteme. Soll zum Test für den K-Fall oder aus Wartungsgründen z.B. nur „Personal“ umgeschaltet werden, so hat man mit VM2000 die Voraussetzung dafür, dass dieses die zwei anderen Anwendungen nicht betrifft, vorausgesetzt, die CPU-Leistung auf dem Standby-Host ist ausreichend dimensioniert.

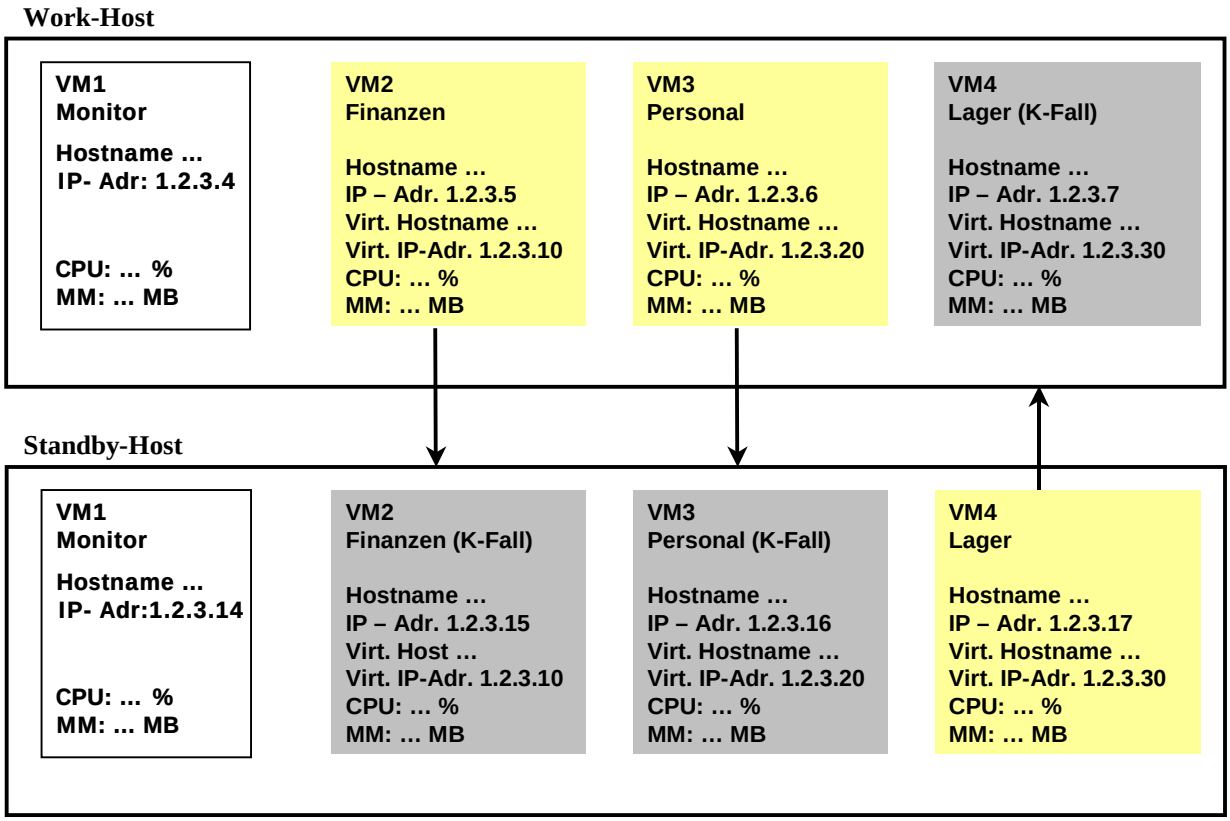


Abb. 4-7 Abschottung der Anwendungen mit VM2000

5 Abläufe beim Failover

Mit Failover bezeichnen wir die Umschaltung der Anwendungen in einem K-Fall vom Produktions-RZ auf das Standby-RZ.

Der Failover ist durchzuführen, wenn ein K-Fall eingetreten ist, d.h. wenn durch eine der in 1.1 genannten Umstände der Betrieb der Anwendungen im Work-RZ nicht mehr möglich ist. Dies ist der Fall, wenn das gesamte RZ ausfällt bzw. nicht mehr nutzbar ist, aber auch, wenn nur ein kritische Daten enthaltendes Symmetrix System im Work-RZ ausgefallen ist, und ein Betrieb des Produktions-Hosts mit den gespiegelten Daten einer Standby-Symmetrix wegen zu geringer Performance (aufgrund einer großen Entfernung) ausscheidet.

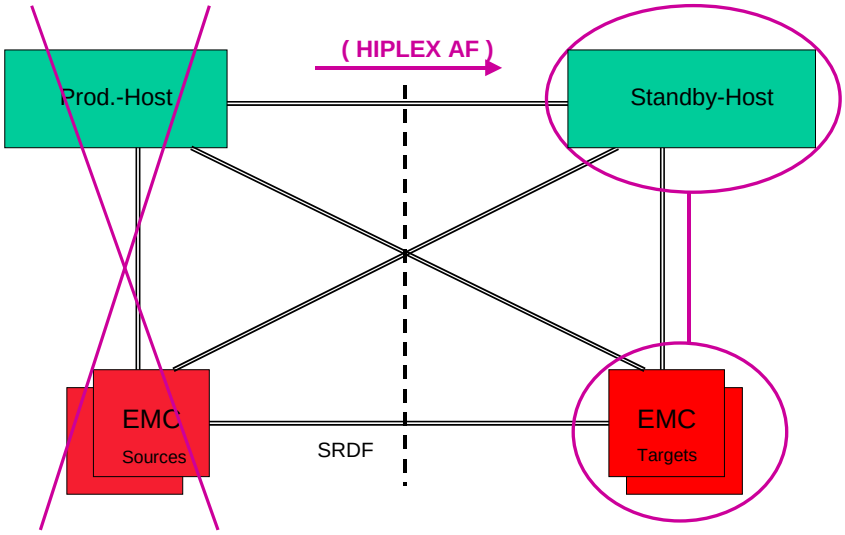


Abb. 5-1 Failover bei Komplettausfall

Wir beschreiben im Folgenden kurz die automatischen Abläufe beim Einsatz von HIPLEX AF für X-Link-Konfigurationen und eine Reihe von Aktionen bei manueller Umschaltung auf das Standby RZ, die im anderen Fall auch von HIPLEX AF ausgeführt werden. Es ist natürlich auch möglich, dass ein von HIPLEX AF gesteuerter Failover „auf Knopfdruck“ manuell angestartet wird, also ein halbautomatischer KS zugrunde liegt.

Die in den folgenden Kapiteln beschriebenen Abläufe beim Failover und Failback sind nicht als vollständige technische Anleitung zu verwenden. Sie sind entstanden aus Projekt- und Testerfahrungen und sollen hier lediglich einen Eindruck des Ablaufs vermitteln.

5.1 Automatischer Failover mit HIPLEX AF

Falls HIPLEX AF im Einsatz ist, was vor allem für die X-Link-Konfigurationen empfehlenswert ist, und falls HIPLEX AF für einen automatischen KS konfiguriert ist, wird das HIPLEX Monitoring am Standby-Host die MSCF-Rekonfiguration einleiten und da sie den Ausfall der Symmetrix aufgrund von SHC-OSD-Meldungen erkennt, den Failover auf die Standby-Symmetrix ausführen. (s. auch Kap. 2.13).

Bei Ausfall des gesamten Work Symmetrix Systems bzw. des gesamten Work-RZ ergibt sich, vereinfacht, folgender automatischer Ablauf:

1. Der SYSRES-Monitor, ein Teil der Überwachung von HIPLEX AF, der die erste Platte des Home Pubsets überwacht, erkennt den Ausfall derselben und terminiert daraufhin das BS2000 im Work-RZ.

2. Das Monitoring von HIPLEX AF am Standby-System erkennt anhand von SHC-OSD-Meldungen den Ausfall der Symmetrix und anhand von MSCF-Meldungen den Ausfall des Work-Systems.

3. Das Monitoring am Standby-System aktiviert die Target Units in der Standby-Symmetrix. Je nach Definition in der Switch-Unit werden entweder alle oder eine vorgegebene Liste von Target Units freigeschaltet.

4. Das Monitoring am Standby-System importiert die Datenpubsets, aktiviert ggf. virtuelle Hosts und startet die Anwendungen mitsamt ihren erforderlichen Ressourcen.

Falls der Ausfall der BCAM-Verbindung und der Ausfall der Work-Symmetrix nicht gleichzeitig, also innerhalb eines durch MSCF-Parameter vorgegebenen Zeitintervalls auftreten, oder falls die „Fail-Rekonfiguration“ bei HIPLEX MSCF **nicht** automatisch anlaufen soll (ausführliche Beschreibung in [4]), muss am Standby-Host die MSCF-Ausfall-Behandlung manuell gestartet werden, da HIPLEX MSCF bei nicht gleichzeitigem Ausfall nicht automatisch von einem Komplettausfall ausgehen darf. Das HIPLEX AF-Monitoring auf dem Standby-Host startet danach die Anwendungen.

5.1.1 Weitere Ausfallszenarien

Außer dem Komplettausfall sind weitere Ausfallsituationen denkbar, die auch automatisch behandelt werden können.

Der in den folgenden Abbildungen gezeigte Ausfall des Work-Hosts wird beim Einsatz von HIPLEX AF i.d.R. wie in Abb. 5-2 dargestellt behandelt: Die Anwendung(en) werden am Standby-Host neu gestartet, jedoch mit den Source Units in der Work-Symmetrix.

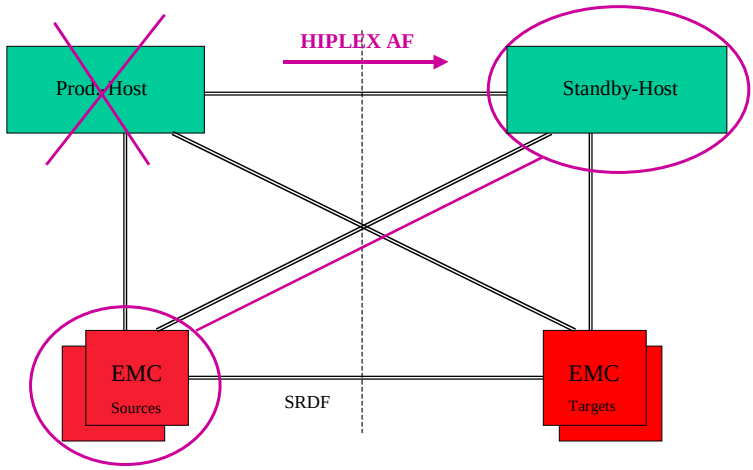


Abb. 5-2 Ausfall des Work-Host → Switchover

Falls es sich um eine Konfiguration über eine größere Entfernung handelt, so dass der IO-Betrieb nicht mehr in der gewünschten Performance laufen würde, muss in diesem Fall ein Failover auf die Standby-Symmetrix ausgeführt werden, wie in Abb. 5-3 gezeigt.

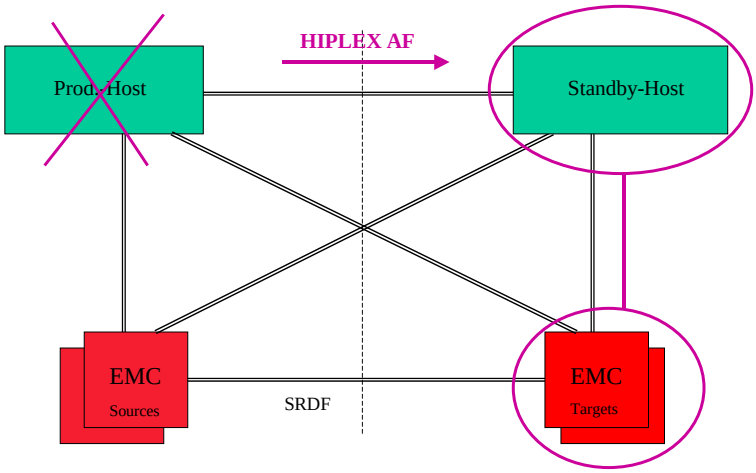


Abb. 5-3 Ausfall des Work-Host → Failover + Switchover

Fällt im Work-RZ nur die Symmetrix aus, wird, wie in Abb. 5-1, die Anwendung auf dem Standby-Host weitergefahren, denn dann wird auch das BS2000 am Work-Host nicht mehr laufen. Beim Einsatz von HIPLEX AF wird das Work-System automatisch terminiert.

Ausfall von Standby-Host oder/ und -Symmetrix:
Falls produktive Anwendungen auch auf dem Standby-Host mit Source Unit in der Standby-Symmetrix laufen, sind aus Sicht dieser Anwendungen die Rollen der beiden Hosts und der beiden Symmetrix Systeme vertauscht, die Abläufe bei Ausfall des Host und/oder der Symmetrix im Standby-Rechenzentrum sind jedoch die gleichen.

Ausfall von Verbindungen:
Wenn, wie schon in Kap. 4.1 beschrieben, die Verbindungen der Einzelkomponenten alle redundant ausgelegt sind und auf unterschiedlichen Wegen verlaufen, es also keine „Single Points of Failure“ gibt, dann wird ein Verbindungsausfall von der jeweiligen Treiber-SW behandelt und stellt somit kein Ausfallszenario dar. Eine derart redundant ausgelegte Verbindungstechnik wird zur Katastrophenvorsorge empfohlen.

5.2 Manueller Failover

Bei KS-Konfigurationen ohne Automatisierung durch HIPLEX AF wird man im K-Fall nach einem Notfallplan oder Notfallhandbuch (s. Kap. 3.1.2) vorgehen, welches natürlich rechenzentrums- und anwendungsspezifisch zu erstellen ist. Es werden daher nur die grundsätzlichen Schritte aufgeführt.

Diese Schritte richten sich aus an der synchronen engen KS-Konfiguration und gelten größtenteils auch für alle in Kapitel 4 beschriebenen Konfigurationen. Eventuelle Abweichungen sind in den konfigurationsspezifischen Kapiteln beschrieben.

Voraussetzung für einen nicht automatisierten Failover ist die sog. K-Fall „Ausrufung“. Es sollte eine oder mehrere Personen (K-Fall-Verantwortliche) geben, die befugt und kompetent sind, zu entscheiden, ob die Notwendigkeit besteht, die Anwendungen auf die Standby-Seite umzuschalten, oder ob es sich nur um einen temporären Ausfall bestimmter HW-Komponenten handelt. Bei dem, was man allgemeinhin als Katastrophe versteht, wie Großbrand oder Überschwemmung, ist diese Entscheidung einfach, aber es kann sich ja z.B. auch nur um einen Stromausfall handeln.

Abfolge der Schritte beim manuellen Failover

1. Abschalten noch laufender Komponenten im Work-RZ: Bevor eine Anwendung im Standby-RZ neu gestartet wird, sollte man sicherstellen, dass im ausgefallenen RZ nicht noch der Work-Host aktiv ist. Falls es also machbar ist, sollten Hosts und möglichst auch die Symmetrix-Systeme im Work-RZ abgeschaltet werden. Falls das nicht möglich sein sollte, weil z.B. das RZ nicht erreichbar ist, müssen die Netzverbindungen zur Außenwelt unterbrochen werden, um zu verhindern, dass der virtuelle Host bzw. die Anwendung nach dem Neustart im Ausweich-RZ im Netz doppelt vorhanden ist.

2. Falls ein MSCF-Verbund eingerichtet ist, muss die MSCF-Rekonfiguration eingeleitet werden, d.h. es werden MSCF-Meldungen so beantwortet, dass die BCAM-Verbindung beendet und nötigenfalls ein Masterwechsel für das Shared Pubset durchgeführt wird.

3. Falls nicht, wie in Kap. 3.3.1 beschrieben, mit virtuellen Hosts und dynamischem Routing gearbeitet wird, müssen die Netzkomponenten umkonfiguriert werden (DNS-Einträge, VLAN-Zuordnung oder Umkonfigurierung des Routing), um Anwenderzugriff auf den Standby-Host zu ermöglichen. Dies ist allerdings unabhängig von den folgenden Schritten.

4. Freischalten der SRDF Target Units am Standby-RZ mit SHC-OSD-Kommandos.

5. Falls ein Cold-Standby Konzept vorliegt, also ggf. auch die Home Pubsets der ausgefallenen Systeme per SRDF gespiegelt wurden, müssen die (Gast-)Systeme jetzt per IPL von den Target Units gestartet werden.

6. Da die Platten der Datenpubsets noch vom Work-System belegt waren, müssen sie per Unlock-Disk Kommando freigegeben werden.

7. Importieren der gespiegelten Datenpubsets von den Target Units.

8. Überprüfung der Anwendungsressourcen: Der Anwendungsbetreuer oder Systemverwalter prüft die Verfügbarkeit und Vollständigkeit der Datenpubsets.

9. Netzumschaltung: Aktivieren der virtuellen Hosts

10. Neustart der Anwendungen: Nach diesen Arbeitsschritten werden die Anwendungen am Standby-Host gestartet.

Die genannten Arbeitsschritte sind natürlich weitgehend automatisierbar und in Prozeduren aufrufbar.

Wie lange kann ein Failover dauern? Pauschal ist diese Frage nicht zu beantworten, bei Tests sind wir bisher stets nach wenigen Minuten bei Punkt 8 angelangt. Wie lange jedoch der Start einer Anwendung und eine Datenbank-Recovery dauert, hängt von der Art der Anwendung und der dafür nötigen Softwarekomponenten sowie von der IO-Last zum Ausfallzeitpunkt ab. Die Zeitdauer bis zur Ausrufung der Katastrophe und der Abschaltung der Ressourcen im Work-RZ ist von den jeweiligen Umständen abhängig und nicht restlos planbar. Die im Kapitel 4 angegebene Größenordnung von 30-45 min Ausfallzeit bietet aber einen guten Anhaltspunkt.

6 Abläufe beim Failback

Im Unterschied zum Failover, der Reaktion auf einen nicht vorhersagbaren K-Fall ist, ist die Rückverlagerung der Anwendungen auf die ursprünglichen Betriebsmittel, also der Failback, planbar! Die durch ihn ausgelöste Anwendungsunterbrechung kann so geplant werden, dass damit verbundene negative Auswirkungen weitgehend vermieden werden können.

Der Failback kann erst dann durchgeführt werden, wenn alle ausgefallenen Komponenten und Verbindungen wieder soweit hergestellt sind, dass die Anwendungen auf das Work-RZ zurückverlagert werden können. Zunächst bringt ein EMC-Techniker die Symmetrix-Steuerungen im Work-RZ wieder in einen brauchbaren SRDF-Zustand: Alle Kanal-Verbindungen und alle Remote Verbindungen werden deaktiviert, alle zugehörigen Kabel wieder angeschlossen, die Symmetrix Systeme hochgefahren, dann werden Oberflächentests für alle Platten durchgeführt. Die SRDF-Verbindungen werden eingerichtet, aber der SRDF-Betrieb noch nicht wieder aufgenommen! Falls das komplette RZ beschädigt war, beginnt der Failback, sobald alle Hosts und zugehörigen Komponenten wieder verkabelt und getestet und alle Netzverbindungen wiederhergestellt und ebenfalls getestet sind. Eine Testversion für jede Produktivanwendung sollte bereits vor dem Failback in dem wiederhergestellten RZ gelaufen sein. Die Dauer des Failbacks hängt wesentlich von der Datenmenge ab, und nimmt in Summe deutlich mehr Zeit in Anspruch als der Failover.

6.1 Failback mit Hilfe von HIPLEX AF

Falls HIPLEX AF im Einsatz ist, übernimmt es das Neuanstarten der Anwendungen und kann zuvor auch den Failback auf die Symmetrix Systeme im Work-RZ ausführen. Der Failback wird nach einem K-Fall i.d.R. nicht vollautomatisch erfolgen, es sind Tätigkeiten an Work- und Standby-Host sowie an den Symmetrix Subsystemen miteinander zu koordinieren. Bei Nutzung von HIPLEX AF wird das System am Standby-Host stets mit einem eigenen Home Pubset laufen (nicht auf Target Units des Work-Systems). Es muss also nicht beendet werden. Trotzdem können Systempubsets auch per SRDF gespiegelt werden, wenn spezielle Daten auf diesen Pubsets im K-Fall benötigt werden.

Abfolge der grundsätzlichen Schritte beim HIPLEX AF gestützten Failback

1. Die Anwendungen und die zugehörigen SW-Ressourcen am Standby-Host werden per HIPLEX AF Kommando beendet. Hiermit werden z.B. auch die Datenpubsets exportiert und ggf. virtuelle Hosts deaktiviert.

2. Die Target Units in der Standby Symmetrix werden wieder für den SRDF Betrieb vorbereitet und damit deaktiviert. Dies kann per HIPLEX AF Failback-Prozedur ausgeführt werden; falls die Symmetrix neu eingerichtet wurde, wird dies der EMC-Techniker tun.

3. Falls die Symmetrix Systeme am Work-RZ neu eingerichtet wurden, muss jetzt ein EMC-Techniker die Resynchronisation von den Target- zu den Source Units anstarten, falls nicht kann dies auch mit der Failback Prozedur von HIPLEX AF durchgeführt werden. Das Work-System darf bei Start der Synchronisierung noch nicht laufen (ggf. bleibt der Host noch abgeschaltet), weil von den Target zu den Source Units synchronisiert wird und gleichzeitig keine Zugriffe auf die Sources erfolgen dürfen.

4. Falls nicht, wie in Kap. 3.3.1 beschrieben, mit virtuellen Hosts und dynamischem Routing gearbeitet wird, müssen jetzt die Netzkomponenten umkonfiguriert werden (DNS-Einträge, VLAN-Zuordnung oder Umkonfigurierung des Routing), um den Anwenderzugriff auf den Work-Host wieder sicherzustellen.

5. Nachdem die Synchronisation für alle Platten angelaufen ist, kann das (Gast-) System am Work-Host hochgefahren werden. In der Regel werden dabei automatisch per Command-File die BCAM-Dateien gestartet, der MSCF-Verbund eingerichtet und das HIPLEX AF Monitoring aktiviert.

6. Die Switch-Units für die Anwendungen können jetzt gestartet werden, diese aktivieren ggf. auch die virtuellen Hosts.

7. Überprüfen und Freigeben der Anwendungen, ggf. werden zuvor Testprozeduren gestartet.

Die Dauer eines solchen Failback hängt im Wesentlichen von der zwischenzeitlich geänderten Datenmenge und von der Bandbreite der SRDF-Verbindung ab. Die Größenordnung der Ausfallzeit bewegt sich nach unseren Erfahrungen im Bereich einer bis weniger Stunden.

6.2 Manueller Failback

Im Folgenden sind die Tätigkeiten für einen manuellen Failback aufgeführt. Diese sind auch z.T. mit Hilfe von Prozeduren automatisierbar. Die Tätigkeiten 1 – 4 sind am Standby-Host auszuführen; falls dieser nur im K-Fall benutzt wird, kann das System danach beendet werden.

Abfolge der grundsätzlichen Schritte beim manuellen Failback

1. Die Anwendung und die zugehörigen SW-Ressourcen am Standby-RZ werden beendet.

2. Die Target Units in der Standby Symmetrix werden wieder für den SRDF Betrieb vorbereitet und damit deaktiviert. Dies erfolgt per SHC-Kommando oder Prozedur am Standby-Host.

3. Falls mit virtuellen Hosts gearbeitet wurde, werden diese im Standby-RZ deaktiviert.

4. Die Datenpubsets werden exportiert. Falls auch das Home Pubset per SRDF gespiegelt wird, muss das System danach beendet werden.

5. Falls nicht, wie in Kap. 3.3.1 beschrieben, mit virtuellen Hosts und dynamischem Routing gearbeitet wird, müssen jetzt die Netzkomponenten umkonfiguriert werden (DNS-Einträge, VLAN-Zuordnung oder Umkonfigurierung des Routing), um den Anwenderzugriff auf den Work-Host wieder sicherzustellen.

6. Falls die Symmetrix Systeme am Work-RZ neu eingerichtet wurden, muss jetzt ein EMC-Techniker die Resynchronisation von den Target zu den Source Units anstarten, falls nicht, kann dies auch per SHC-OSD-Kommandos oder Prozeduren erfolgen.

7. Nachdem die Synchronisation für alle Platten angelaufen ist, können das (Gast-) System am Work-Host hochgefahren und die BCAM-Dateien gestartet werden.

8. Falls ein HIPLEX-MSCF-Verbund genutzt wird, kann man diesen jetzt wieder aufbauen.

9. Die Datenpubsets werden zugeschaltet. Falls die Platten noch vom SHC-OSD des Standby-Systems belegt sind (durch die vorhergehende Aktion 2), muss ggf. ein Unlock-Kommando auf die Source Units abgesetzt werden, bevor die Pubsets importiert werden können.

10. Ein Anwendungsbetreuer führt eine Standard-Überprüfung des Zustands der Anwendungen und der zugehörigen Ressourcen wie Datenbanken und Dateien (Logging, ggf. manuelles Schließen von Dateien erforderlich) durch.

11. Falls mit virtuellen Hosts gearbeitet wird, werden diese im Work-RZ wieder aktiviert.

12. Die Anwendungen können am Work-Host gestartet werden

Zum vorliegenden Dokument gibt es ein Nachfolge-Papier [10], das analog zu den hier beschriebenen Katastrophenschutz-Konzepten speziell einen Überblick über die Möglichkeiten beim Aufbau von Katastrophenschutz-Konfigurationen mit den Business Servern der SX-Linie gibt, wobei insbesondere auf die neuen Aspekte eingegangen wird, die sich durch die SX-HW-Plattform, durch den möglichen parallelen Einsatz der zwei Betriebssysteme BS2000 und Solaris, und die Storage Systeme vom Typ FibreCAT ergeben.

7 Literatur und Online-Verweise

[1] Benutzerhandbuch „SHC-OSD V5.0 / SCCA-BS2 V1.0 (BS2000/OSD)“
U41000-J-Z125-5
http://manuals.ts.fujitsu.com/ (mainframes → SHC-OSD)

[2] Benutzerhandbuch „DAB V9.0 Disk Access Buffer“
U2431-J-Z125-14
http://manuals.ts.fujitsu.com/ (mainframes → DAB)

[3] Benutzerhandbuch „HIPLEX AF V3.2 - Hochverfügbarkeit von Anwendungen in BS2000/OSD“
U24401-J-Z125-4
http://manuals.ts.fujitsu.com/ (mainframes → HIPLEX AF)

[4] Benutzerhandbuch „HIPLEX MSCF V3.0 – BS2000 Rechner im Verbund“
U3615-J-Z125-8
http://manuals.ts.fujitsu.com/ (mainframes → HIPLEX MSCF)

[5] Fujitsu White Paper: HIPLEX - Der BS2000/OSD-Cluster
https://sp.ts.fujitsu.com/dmsp/docs/wp_hiplex_de.pdf

[6] IT-Grundschutzhandbuch / Standard-Sicherheitsmaßnahmen
http://www.bsi.bund.de/gshb
Bundesamt für Sicherheit in der Informationstechnik

[7] Wavelength Division Multiplex
Sebastian Groller, Ausarbeitung für Seminar Rechner- und Betriebssysteme
http://www.uni-weimar.de/~grolla/docs/wdm/index.html

[8] IBM System Storage Business Continuity: Part 1 Planning Guide
IBM Redbook SG24-6547-03
http://www.redbooks.ibm.com/

[9] EMC Support Matrix
http://www.emc.com/interoperability/index.jsp

[10] Fujitsu White Paper: Katastrophenschutz-Konzepte für SX-Server
https://sp.ts.fujitsu.com/dmsp/docs/wp_katastro-sx_de.pdf

Alle Rechte vorbehalten, insbesondere gewerbliche Schutzrechte. Änderung von technischen Daten sowie Lieferbarkeit vorbehalten. Haftung oder Garantie für Vollständigkeit, Aktualität und Richtigkeit der angegebenen Daten und Abbildungen ausgeschlossen. Wiedergegebene Bezeichnungen können Marken und/oder Urheberrechte sein, deren Benutzung durch Dritte für eigene Zwecke die Rechte der Inhaber verletzen kann.
Weitere Einzelheiten unter ts.fujitsu.com/terms_of_use.html

Herausgegeben durch:

Margret Germann
Telefon: ++49 (0)89 62060-1975
Fax: ++49 (0)89 62060-329-1975
Margret.Germann@ts.fujitsu.com
de.ts.fujitsu.com

Partner login
partners.ts.fujitsu.com

Copyright © Fujitsu Technology Solutions GmbH 2009